UNIVERSITÀ DI PADOVA FACOLTÀ DI INGEGNERIA

Dipartimento di Ingegneria dell'Informazione

Scuola di Dottorato di Ricerca in Ingegneria dell'Informazione
Indirizzo: Bioingegneria

XXIII CICLO

MASS SPECTROMETRY-BASED PROTEOMICS: A 3D APPROACH TO DATA HANDLING AND QUANTIFICATION

Direttore della Scuola: Ch.mo Prof. Matteo Bertocco

Supervisore: Dott.ssa Barbara Di Camillo

Dottoranda: Sara Nasso

ABSTRACT

This thesis describes the Ph.D. research project in Bioengineering for Computational Proteomics carried out during the last three years (January 2008 - January 2011). Activities focused on design and development of methods for the analysis of Quantitative Mass Spectrometry-based Proteomics data.

The Introduction briefly elucidates the main themes developed in the thesis and how the work was schemed. It reviews the computational issues associated to both data handling and quantification, and introduces the solutions proposed in the following.

The first two chapters are introductory to the Proteomics and Mass Spectrometry field. The objective is to provide the reader with the information needed to understand Quantitative Mass Spectrometry-based Proteomics. In particular, Chapter 1 explains how proteomics was born, as the –omics science of proteins. Then proteomics main applications and goals are illustrated, which are ranging from clinics and pharmaceutics to systems biology. Chapter 2 shows the main technologies and instrumentation exploited in Mass Spectrometry-based proteomics. The most common experimental setups are reported: among them, the Liquid Chromatography-Mass Spectrometry (LC-MS) technique is thoroughly explained since it is the principal technique for Quantitative Mass Spectrometry-based Proteomics.

The third Chapter presents the main concepts necessary to introduce the reader to the main topic of the PhD research Project, that is the development of bioinformatics tools for the handling and quantification of Mass Spectrometry-based Quantitative Proteomics data, focusing on LC-MS quantitative data and their analysis. Indeed, LC-MS data are highly informative for quantification aims, but challenging to parse. Data features that were pivotal for the design of the proposed solutions (i.e., the 3D structure of LC-MS data and the high quality profile acquisition) are highlighted.

In the fourth Chapter, the state of art both for data handling and quantification is described and available standard data formats and software are illustrated as well as related open challenges.

In Chapter 5, the dataset used to carry out the analyses is technically described. It consists of LC-MS data from a labeled controlled mixture of proteins with known quantification ratios, acquired in profile acquisition mode and in triplicates.

In particular, this thesis presents 2 software solutions to address the handling and quantification of Quantitative Mass Spectrometry-based Proteomics data: mzRTree and 3DSpectra, respectively.

Chapter 6 presents the solution proposed for the data handling issue. The proposal is a scalable 2D indexing approach implemented through an R-tree-based data structure, called mzRTree, that relies on a sparse matrix representation of the dataset, which is appropriate for LC-MS data, and more in generally for MS-based proteomics data. mzRTree allows efficient data access, storage and enables a computationally sustainable analysis of profile MS data.

Regarding the quantification, which is one of the most relevant problem in mass spectrometry-based proteomics, Chapter 7 illustrates the solution proposed for the quantification problem: 3DSpectra. It is an innovative quantification algorithm for LC-MS labeled profile data exploiting both the 3-dimensionality of data and the profile acquisition. 3DSpectra fits on peptide data the 3D isotopic distribution model shaped by a Gaussian Mixture Model including a noise component, using the Expectation-Maximization approach. This model enables the software to both recognize the borders of the 3D isotopic distribution and reject noise. 3DSpectra is a reliable and accurate quantification strategy for labeled LC-MS data, providing significantly wide and reproducible proteome coverage.

In the conclusion section of this thesis future and ongoing research work, regarding further development of both the mzRTree data structure and 3DSpectra quantification software, are discussed.

SOMMARIO

La presente tesi descrive il progetto di ricerca in Bioingegneria per la Proteomica Computazionale svolto durante i tre anni di dottorato (Gennaio 2008 - Gennaio 2011). L'attività di ricerca è stata incentrata sulla progettazione e lo sviluppo di metodi per l'analisi di dati di Proteomica basata su Spettrometria di Massa.

Nell'introduzione si illustrano brevemente i temi principali trattati nella tesi, fornendo così lo schema del lavoro svolto. Si considerano quindi i 2 problemi principali associati all'analisi dati, cioè la gestione e quantificazione dei dati, e vengono presentate le soluzioni descritte nel prosieguo.

I primi due capitoli sono introduttivi al settore della Proteomica e della Spettrometria di Massa. L'obiettivo è fornire al lettore tutte le informazioni necessarie per meglio comprendere la Proteomica Quantitativa basata su Spettrometria di Massa. Il Capitolo 1 spiega in che modo sia nata la Proteomica, ossia come il complemento proteico del genoma. Dopodiché, si espongono le principali applicazioni legate alla Proteomica e i suoi obiettivi, spaziando dagli aspetti clinici, alla farmaceutica, fino alla biologia dei sistemi. Il secondo Capitolo invece è legato agli aspetti tecnici e mostra le principali tecnologie e strumentazioni usate in Proteomica basata su Spettrometria di Massa. I setup sperimentali più comuni sono quindi illustrati e, tra tutti, ci si focalizza in particolare sulla Spettrometria di Massa abbinata a Cromatografia Liquida (LC-MS), che è la principale tecnica per esperimenti di Proteomica Quantitativa basata su Spettrometria di Massa.

Il terzo Capitolo presenta i concetti fondamentali necessari per introdurre il lettore al tema principale del progetto di ricerca di Dottorato, ossia lo sviluppo di metodi bioinformatici per la gestione e la quantificazione di dati di Proteomica Quantitativa basata su Spettrometria di Massa, in particolare per l'analisi di dati quantitativi di LC-MS. Infatti, i dati di LC-MS hanno un alto contenuto informativo per scopi quantitativi, però sono estremamente problematici da analizzare. Sono quindi riassunti i setup sperimentali per la Proteomica Quantitativa basata su LC-MS così come le caratteristiche dei dati che sono state determinanti per lo sviluppo delle soluzioni proposte (ossia la struttura 3D dei dati LC-MS e l'alto contenuto informativo dei dati profile).

Nel quarto Capitolo vengono descritti lo stato dell'arte, sia per la gestione che la quantificazione dei dati, e i relativi problemi aperti, che verranno trattati nei capitoli seguenti dove si propongono possibili soluzioni.

Il Capitolo 5 è interamente dedicato alla descrizione tecnica dei dati utilizzati per validare le metodologie proposte. Si tratta di dati LC-MS generati da una mistura di proteine tracciate ed a rapporti di quantificazione note. Di ogni esperimento sono disponibili tre repliche.

In particolare, questa tesi presenta 2 software per la gestione e la quantificazione di dati di Proteomica Quantitativa basata su Spettrometria di Massa.

Il Capitolo 6 presenta la soluzione proposta per risolvere i problemi di gestione dati. Si tratta di un approccio di indicizzazione 2D scalabile che è stato implementato tramite una struttura dati basata sull'R-tree, chiamata mzRTree, e si basa sulla rappresentazione del dataset come matrice sparsa, che ben si adatta a dati di LC-MS e più in generale di Spettrometria di Massa. Nello specifico, mzRTree consente di accedere e memorizzare efficientemente i dati, rendendo così possibile un'analisi computazionalmente sostenibile di dati profile.

Per quel che concerne la quantificazione, il Capitolo 7 illustra la soluzione proposta per il problema della quantificazione, 3DSpectra, un innovativo metodo di quantificazione che sfrutta sia la 3-dimensionalità dei dati LC-MS, sia l'alto contenuto informativo dei dati profile. 3DSpectra applica infatti un approccio 3D al riconoscimento della distribuzione isotopica del peptide da quantificare basato sul fit tramite l'algoritmo Expectation-Maximization di un Modello 3D a Mistura di Gaussiane. Tale modello consente di identificare i bordi del segnale da quantificare e di rigettare il rumore presente. 3DSpectra incorpora un'affidabile ed accurata strategia di quantificazione per dati LC-MS tracciati e acquisiti in modalità profile. Soprattutto, 3DSpectra offre, a livello di quantificazione, un'estesa e riproducibile copertura del proteoma.

Nella sezione conclusiva della tesi si discute il lavoro futuro e in corso, che riguarda essenzialmente ulteriori sviluppi sia della struttura dati, mzRTree, che del software di quantificazione, 3DSpectra.

INTRODUCTION

Mass spectrometry-based proteomics plays an ever-increasing role in different biological and medical fields, but, as an emerging field, it still requires reliable tools for the storage, exchange and analysis of experimental data. Over the last years, a wide range of techniques have become available, which can generate a huge quantity of data potentially able to address relevant questions, e.g., to identify proteins in a biological sample (qualitative approach), to quantify their concentration (quantitative approach), to monitor post-translational modifications, to measure individual protein turnover, to infer on interactions with other proteins, transcripts, drugs or molecules. The improved proteomics technologies enable researchers to address fundamental biological problems in a systems biology context but, without efficient bioinformatics tools, high-throughput proteomics data handling and analysis are difficult and error-prone. Thus, a major challenge facing proteomic research is how to manage the overwhelming amount of data in order to extract the qualitative and/or quantitative information on proteome and still to keep down computational costs both for data handling and processing. This holds especially for quantitative proteomics, since, in order to achieve reliable quantifications, it needs highly informative but challenging to parse profile data, such as profile Liquid Chromatography-Mass Spectrometry (LC-MS) datasets, which are considered the only data source rich enough to perform a meaningful data analysis.

DATA HANDLING

Data hostage held by different instrument proprietary formats slows down the evolution of proteomics, mainly because comparisons among different experiments or analytical methods often turn out to be unfeasible. In order to facilitate data exchange and management, the Human Proteome Organization (HUPO) established the Proteomics Standards Initiative (PSI). HUPO-PSI released the Minimum Information About a Proteomics Experiment (MIAPE) reporting guidelines and proposed mzData which, as mzXML, is an eXtensible Markup Language (XML) based data format, developed to uniform the community data. Recently, merging the best features of each of these formats, the HUPO introduced mzML as a unique data format. XML-based data formats are characterized by intuitive language and a standardized structure. At the state of art, the adoption of these formats is widespread among the proteomics research groups, also thanks to the extensive support of instrument and database searching vendors, and the availability of converters from proprietary data formats. In spite of their success, the currently adopted formats suffer from some limitations: the impossibility to store raw data; the lack of information on the experimental design, necessary for regulatory submission; the lack of scalability on data size, which is a bottleneck for the analysis of profile data. Above all, the 1-dimensional (1D) data indexing provided by these formats considerably penalizes the analysis of datasets embodying an inherent 2-dimensional (2D) indexing structure, such as 3D LC-MS data. LC-MS provides intensity data on a 2D (t, m/z) domain, since LC separates proteins along retention time dimension (temporal index) based on their chemical-physical properties, while MS separates proteins based on their mass over charge (m/z index) ratios. MS experiments usually have a "temporal" index related to the experimental time at which the MS acquisition takes place (e.g., a scan in mzML format). Thus, we can conceptually view an LC-MS (or, more generally, MS) dataset as a matrix, where the rows are indexed by retention times (scan if MS), the columns by m/z values, and the indexed values are intensities. Hence, a generic entry can be denoted as (rt, mz; I), where rt and mz are the row and column indices, and I is the intensity value. Therefore, MS data can be accessed by means of either an m/z range, or a temporal range, or a combination of them, defining different range queries. On LC-MS data, these accesses provide respectively chromatograms, spectra, and peptide data, whereas on generic MS data, they provide a set of sub-spectra belonging to the specified range. An elevated number of range queries are required during data analysis, thus optimizing them would significantly improve computational performance. Depending on the downstream analysis, data can be retrieved as a 2D or a 3D signal. Most research groups develop, often in a sub-optimal way, intermediate data structures optimized for accesses on a privileged dimension: the lack of a gold standard for data analysis delayed the development of a standard data format optimized for computation, indeed. For instance, accredited software packages like Maspectras or MapQuant make use of the method-specific intermediate data structures Chrom and OpenRaw, respectively: the former is optimized for a chromatogram based access, the latter for a spectra based access.

During PhD research activities, a novel data structure, called mzRTree, was developed to efficiently access high-throughput LC-MS profile datasets. It combines a hybrid sparse/dense matrix representation of the data and a scalable index based on the R-tree. In this thesis, it is experimentally shown that mzRTree supports efficiently both 1D and 2D data accesses. In particular, mzRTree significantly outperforms Chrom and OpenRaw on small and large peptide range queries, yielding in some cases orders of magnitude improvements. Furthermore, it still ensures best performance on the accesses for which the other data structures are optimized, i.e., chromatograms for Chrom and spectra for OpenRaw. The experiments also provide evidence that mzRTree is more space efficient than Chrom and OpenRaw, and exhibits good scalability on increasing dataset densities. Therefore, mzRTree is suitable for high density/large size proteomics data, such as profile data, considered as the most informative and hence the most suitable to tackle quantification aims. At present, profile data size reaches several GBs, and it is expected to further raise, as far as instrument accuracy and resolution increase: even a narrow range of m/z values can be challenging to manage when analyzing these data. Thus, the adoption of mzRTree for data storage could make profile data accessible for analysis purposes: it prevents out-of-memory errors, often occurring with huge profile proteomics datasets, and reduces the need for (and the costs of) extraordinary computational infrastructures and their management. Actually, profile data are often the only data source rich enough to carry out a meaningful analysis, e.g., in quantitative proteomics based on stable isotope labeling. However, costs involved with profile data handling often outweigh their benefits. mzRTree could revert this relationship.

QUANTIFICATION

During the last decade many research groups developed quantification software to analyze their own data: most of this software accepts few data formats often generated by a single instrument, data should be produced under a particular experimental workflow, and their quantification performance has been poorly assessed. Conversely, some of them, developed for a widespread use, such as the freely available ASAPRatio or the licensed Mascot Distiller, showed good performance and are commonly used among proteomic research laboratories. At the state of art, quantitative LC-MS data have usually been analyzed throughout a 2D approach: all intensities belonging to a defined m/z range related to a peptide were integrated to get a unique chromatogram of the elution profile. Such an approach, reducing a 3D signal to a 2D signal does not involve just a complexity reduction, but, above all, the loss of the LC-MS instrumentation resolving power and therefore the waste of meaningful information, causing neighboring peaks to overlap on the LC dimension: as a result we can achieve unreliable quantifications. Hence the need to develop a 3D approach. In fact, the 2D-LC-MS separation (t, m/z), raising resolving power, minimizes the overlapping of neighbouring peptides, while the profile acquisition mode enhances the signal informative content, consequently the quantification gets more accurate.

Therefore, during this PhD research project, both data features were exploited and 3DSpectra, an innovative quantification software for LC-MS labeled profile data, was developed under MATLAB environment. 3DSpectra features an optimized profile data handling, by means of mzRTree, and a hybrid 2D and 3D data analysis approach, where a 2D signal processing on both chromatograms and spectrograms is coupled to a 3D peaks borders recognition step. 3DSpectra makes use of a priori information, provided by search engines, to quantify identified peptides, whose metadata are stored in a peptide library. It fits on peptide data the isotopic distribution shaped by a 2D Gaussian Mixture Model (GMM) including a noise component, using the Expectation-Maximization (EM) approach, in order to statistically define its boundaries. Data outlying the borders or belonging to the noise component are discarded from subsequent analysis. After signal processing, information gathered from metadata is used to weight the isotopic peaks contribution to the volume under the curve (VUC) of the isotopic distribution. The quantification is computed as the ratio of the peptide VUC to its isotopic partner VUC. 3DSpectra performance has been assessed employing real profile data from a controlled mixture of labeled proteins mixed at different ratios in triplicates and acquired in enhanced profile mode. Quantification performance on this dataset has already been published showing that ASAPRatio (MASPECTRAS implementation) reaches the best performance compared to MSQuant and PepQuan. Consequently we compared 3DSpectra only to ASAPRatio (MASPECTRAS implementation). The comparison focused on the following quantification quality parameters: accuracy, precision, efficiency, reproducibility and reliability. In order to make the comparison as fair as possible both methods have been run starting from the same peptide identifications. 3DSpectra quantifies, on differentially expressed ratios, 2 to 4 times more peptides than ASAPRatio, resulting in a 100% to 300% gain in quantification efficiency (i.e., the number of quantified peptides). Furthermore, the wider proteome coverage here comes with no tradeoff: 3DSpectra, indeed, reaches the same performance as ASAPRatio for quantification accuracy, precision and reliability. In fact, quantifications provided by 3DSpectra and ASAPRatio for every ratio are not statistically different (Kolmogorov-Smirnov test). The much wider peptidome coverage coupled to the same quantification accuracy and precision of ASAPRatio, as provided by 3DSpectra, could be crucial for biomarkers discovery studies. Likewise, the quantification reproducibility, e.g., the ability to reliably quantify the same peptide across experimental replicates, could be pivotal as well. In fact, it could help classification algorithms in distinguishing differentially expressed peptides between control and unhealthy samples, especially when several samples are available per every class. 3DSpectra achieves a significantly higher reproducibility of its peptide quantifications across experimental replicates, quantifying 30% more peptide occurrences than ASAPRatio does, still ensuring the same quantification accuracy and precision. Moreover, 3DSpectra Deming regressions between light and heavy volumes showed on mean higher linearity (Pearson correlation coefficient) than ASAPRatio and comparable Root Mean Squared Error on the same peptides, hence the two methods feature the same quantification reliability. In conclusion, 3DSpectra, compared to ASAPRatio, provides a reliable quantification strategy and a wider and more reproducible proteome coverage at the level of peptide quantification.

In the next chapters we will go deeply through all the computational and methodological issues here introduced, which have been studied during this Ph.D. research project.

1 PROTEOMICS

The term “proteome” refers to the collection of proteins within a cell, tissue, or entire organism and was first coined to describe large-scale protein identification and amino acid analysis: it represents the entire complement of proteins expressed by a cell under a specific set of conditions at a specific time.

Proteomics is the large-scale study of proteins focused on their structures, functions and regulatory physiological pathways. For physiologists and physicians interested in the regulation of bodily functions, an understanding of genes and their products is crucial to unrevealing the underlying mechanisms of disease. Comprehending the regulation of both normal physiology and pathology requires an investigation of genes, gene transcripts, proteins, and metabolites, which have been termed the genome, transcriptome, proteome, and metabolome, respectively. Perhaps the most important step in the expression of a gene occurs at the level of protein synthesis, since the protein product of a gene is what will ultimately be responsible for most biological functions.

In order to fully understand proteomics, one must first understand what proteins are. A protein is a macromolecule that consists of a long chain of amino acids. This amino acid chain is translated according to RNA sequence that, in turn, is transcribed from DNA. This progression from DNA to RNA and then RNA to protein is often known as the central dogma of molecular biology.

There are four “levels” of protein structure. The first, called primary structure, is the sequence of amino acids that makes up a protein. Twenty different amino acids make up the standard protein alphabet utilized by organisms. Secondary structure includes local interactions between groups of amino acids, forming structures such as α-helices and β-sheets (see Figure 1-1 where they are respectively represented in red and blue).

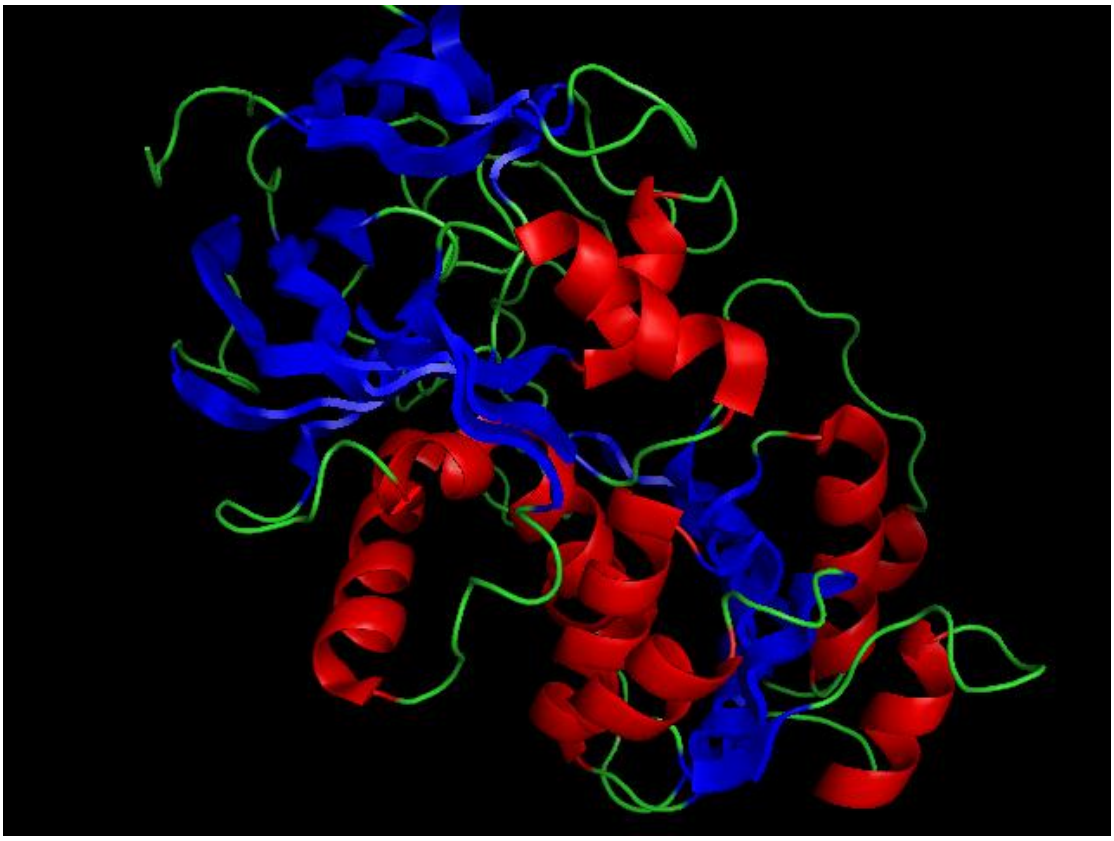


FIGURE 1-1 THE FEATURES SHOWN IN RED REPRESENT ALPHA-HELICES, AND BLUE REPRESENTS BETA-SHEETS.

Tertiary structure is the overall three dimensional conformation of a protein, which can include interaction between secondary structure units. Often, the active form of a protein will actually consist of multiple smaller protein units which combine to yield quaternary structure. Ultimately, the order of amino acids and interactions between them determines the three-dimensional structure the protein will eventually take on. This 3-D structure determines the function of the protein. The process of going from primary structure to tertiary or quaternary structure is often called folding and docking.

Besides protein identification, proteomics also encompasses the regulation of protein synthesis at the translational level, the study of factors regulating the folding of peptides, and interactions among proteins. The complexity of proteomics is further magnified by the fact that protein expression is tissue specific, and its function is modulated by a variety of factors: it varies among different tissues as well as different physiological conditions, such as age, sex, fasting and feeding, changes in diet, physical activity, medications, pregnancy, disease status, etc.

Understanding how multitudes of proteins change under these conditions will be a great challenge to physiologists and clinicians.

Proteomics has recently demonstrated its utility in understanding cellular processes on the molecular level as a component of systems biology approaches and for identifying potential biomarkers of various disease states [1,2]. The large amount of data generated by utilizing high efficiency (e.g. chromatographic) separations coupled to high mass accuracy mass spectrometry for high-throughput proteomics analyses presents challenges related to data processing, analysis, and display. Exploration of a proteome depends not only on establishing robust high-throughput methods for sample analysis, but also on finding solutions to the subsequent challenge of extracting the desired information from the vast quantities of data that are commonly produced in both systems biology and candidate biomarker discovery efforts. Therefore the state of bioinformatics is critical for interpretation of the vast amount of information emerging from proteomic research. To unravel the underlying systems biology mechanism there is a compelling need for greater integration of proteomic research with genomic, metabolic, and functional studies. Actually an omics-integration, able to figure out unknown biological inferences, is what systems biology is trying to realize. In Figure 1-2, the systems biology paradigm is represented. Here, cells are subjected to specific (e.g. genetic or pharmacological) perturbations within the space of the system studied and the effects of the perturbations on the cells are recorded using systematic genomic and proteomic methods of analysis. Proteomic data that are particularly informative include quantitative protein profiles, profiles of regulatory modifications and protein interaction networks. The data are integrated and reconciled with prior models describing the studied system and discrepancies between the observed data and the model are used to design new perturbations, which are analyzed by means of systematic measurements. The process is repeated iteratively until model and observed data converge. A systems level understanding of organisms is likely to increasingly impact biomedical research, drug discovery, nutrition science, and clinical practices [3]. The ability to broadly measure biological macromolecules, especially proteins, in a high-throughput manner is essential for delineating complex cellular networks and pathways and the response of these pathways to biological stressors.

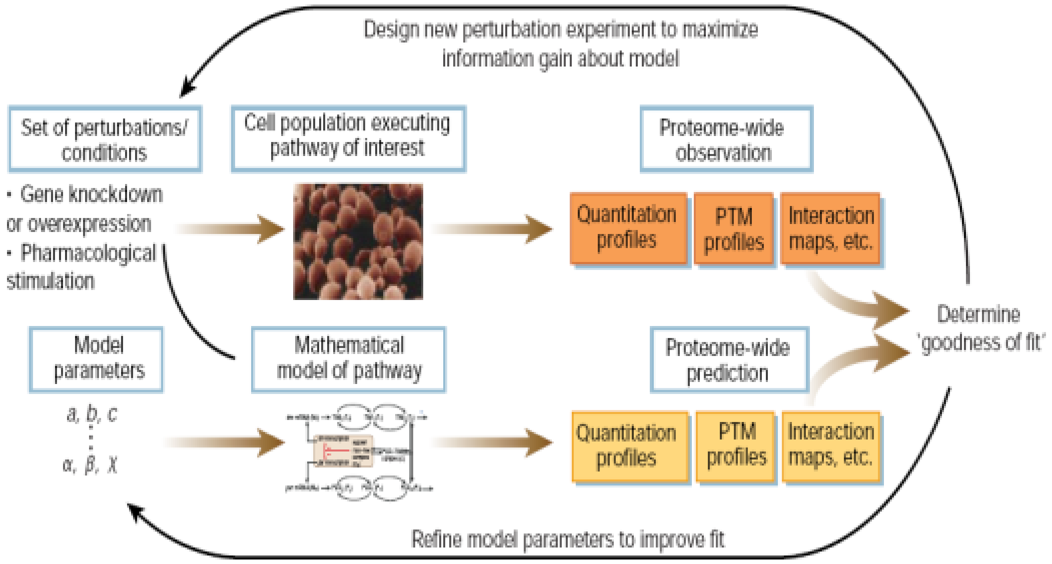

Design new perturbation experiment to maximize information gain about model
Set of perturbations/ conditions
Cell population executing pathway of interest
Proteome-wide observation
• Gene knockdown or overexpression
• Pharmacological stimulation
Quantitation profiles
PTM profiles
Interaction maps, etc.
Determine 'goodness of fit'
Model parameters
Mathematical model of pathway
Proteome-wide prediction
a, b, c
α, β, χ
Quantitation profiles
PTM profiles
Interaction maps, etc.
Refine model parameters to improve fit

FIGURE 1-2 SCHEMATIC REPRESENTATION OF THE SYSTEMS BIOLOGY PARADIGM.

While the genome of an organism may be considered static, the expression of that genome as gene products (i.e. proteome) is constantly changing due to the influence of environmental and physiological conditions. For example, both mRNAs and proteins can be expressed, modified, and degraded at substantially different rates. Thus, measuring the changes in protein expression in response to cellular stressors provides important information on the underlying processes. This information can lead to a better understanding of disease processes in humans, which can aid in the development of novel drug therapies. In this regard there is broad interest in identifying proteins as potential biomarkers for a wide range of diagnostic and clinical applications.

In the following of this chapter several cutting edge applications of proteomics research are illustrated, such as the identification of new proteins, the discovery of biomarkers, the analysis of post-translational modifications (PTMs), proteins turnover, protein-protein interaction, drugs discovery and the role of proteomics in the systems biology field.

1.1 A BRIEF HISTORY

The term proteome was coined by the Macquarie University PhD candidate Mark Wilkins first in 1994 in the symposium: "2D Electrophoresis: from protein maps to genomes" in Siena, Italy. The term arose out of Wilkins's search for an alternative to the phrase "the protein complement of the genome". Actually the term proteome is a blend of proteins and genome and Wilkins used it to describe the entire complement of proteins expressed by a genome, cell, tissue or organism. Subsequently this term has been specified to contain all the expressed proteins at a given time point under defined conditions. The word "proteome" is now firmly established in mainstream scientific language, and while Wilkins and co-workers are rightly credited for formalising "proteomics" as a unique discipline, the origins of proteomics can be traced back to the 1970-80s.

Proteomics has its roots in analytical biochemical techniques used for protein separation. The first high resolution protein separations were achieved by two-dimensional gel electrophoresis (2DE) in 1975, long before global differential analysis of mRNA expression was possible. Proteomic pioneers such as Leigh Anderson saw the potential of 2-D gels in the late 70s, as a mechanism to conduct proteomic studies of blood proteins and leukocytes. The first computerised 2-D gel image analysis platform was developed to quantitate changes in 2-D gel protein spot levels. While the separation of hundreds of proteins using 2-D gels was welcomed, and changes in protein abundance between samples could be quantitated, frustration also grew with the lack of useful tools to identify proteins of interest. Furthermore, 2-D gel reproducibility hindered the expansion of the technique until the introduction of immobilized pH gradients (IPGs) in 1982, and the much improved second generation IPGs in the late 80s. This coincided with the development of mass spectrometry ionization techniques for peptides, allowing protein identification and characterisation on a large scale. Meanwhile, since the 1970s, it has been suggested to build up protein databases and many of the analytical methods nowadays used for the analysis of genomics and proteomics data were born, like reverse strategies based on subtractive pattern analysis, multivariate statistics, clustering algorithms. Unfortunately they couldn't implement those concepts basically because 2DE was just a qualitative technique.

However, it was not until the mid-90s that mass spectrometry (e.g. MALDI-MS, ESI-MS/MS that we will discuss later on) became a mainstream technique for protein identification. Finally protein chemists have been able to create sequence databases and thus database search tool. In the following years, with the decoding of several genomes, the size of translated protein databases ballooned. In the meantime the gel-independent approach to proteomics (i.e. LC-MS/MS) took place thanks to its ability to handle extremely complex peptide mixtures and to facilitate high-throughput experiments (see Figure 1-3) combining very high resolution and high efficiency separations with very high accuracy and high-resolution mass spectrometry.

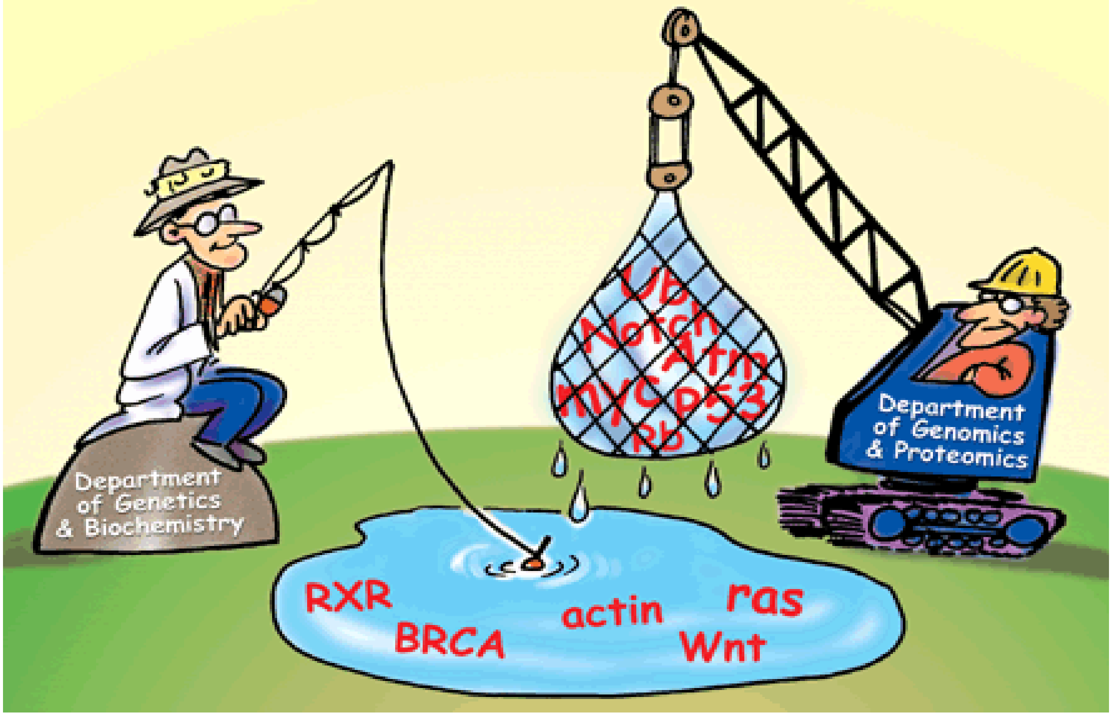

Department of Genetics & Biochemistry
Ubi
Notch
Atm
myc
p53
Rb
Department of Genomics & Proteomics
RXR
BRCA
actin
ras
Wnt

FIGURE 1-3 A NICE PICTURE INTUITIVELY EXPLAINING THE DIFFERENCE BETWEEN THE CLASSICAL CHEMIST APPROACH AND THE HIGH-THROUGHPUT PROTEOMICS ONE.

Significant technological advances in proteomics approaches and instrumentation, as well as in related bioinformatics data analysis, have been achieved over the past decade (see Figure 1-4). In proteomic labs it is now possible to robustly separate complex protein mixtures with high resolution, extract the proteins of interest and interrogate them with mass spectrometry, and then ultimately search protein databases using mass spectral data to identify proteins with high confidence.

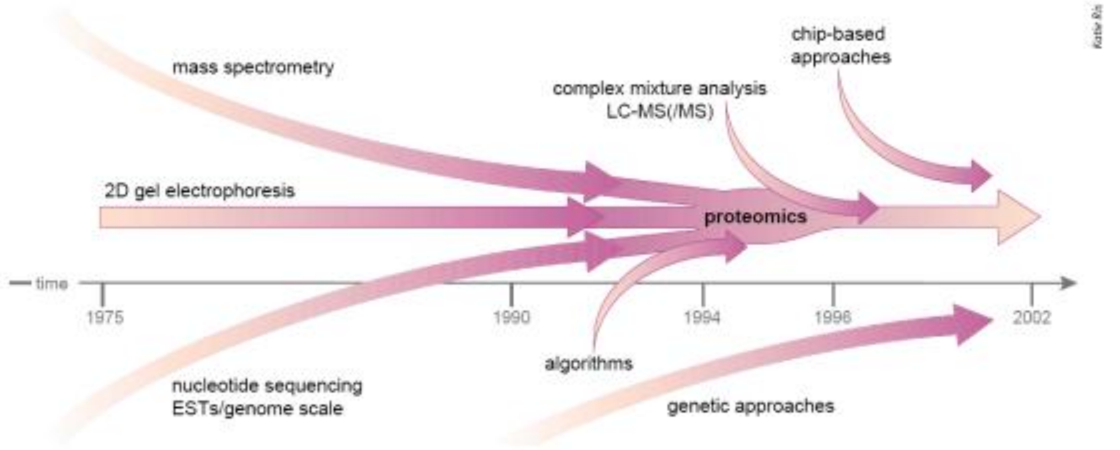


FIGURE 1-4 THE PROTEOMICS TIMELINE. IT DESCRIBES THE ONCOMING OF DIFFERENT TECHNOLOGIES AND RESOURCES, SUCH AS BIOINFORMATICS, MASS SPECTROMETRY AND THE GENOME SEQUENCING, TO THE PROTEOMICS FIELD.

1.2 APPLICATIONS & GOALS

Proteomics has a wide range of applications and they are all focused on the biomedical research field because understanding the proteome, the structure and function of each protein and the complexities of protein-protein interactions will be critical for developing the most effective diagnostic techniques and disease treatments in the future.

1.2.1 NEW PROTEINS IDENTIFICATION

Proteomics is often considered the next step in the study of biological systems, after genomics. It is much more complicated than genomics, mainly because while an organism's genome is rather constant, a proteome differs from cell to cell and constantly changes through its biochemical interactions with genome and environment. One organism has radically different protein expression in different parts of its body, different stages of its life cycle and different environmental conditions. Another major difficulty is the complexity of proteins relative to nucleic acids: in human there are about 25.000 identified genes but an estimated number of more than 500.000 proteins, mostly unknown, that are derived from these genes. Increased complication derives from mechanisms such as alternative splicing, protein modification (glycosylation, phosphorylation) and protein degradation. These processes modify the proteome during the instrumental acquisition time: every protein concentration is constantly modulated by the balance of different appearance/disappearance rates due to the above mentioned processes. The number of proteins in any tissue is likely to be in the tens of thousands, and the expression levels of these proteins span at least six orders of magnitude. In such a complexity it is evident that we know just a smaller subsets of the existing proteins and moreover we identified the most expressed (i.e., concentrated), that are, almost ever, the less informative. Thus a consistent effort in proteomics research nowadays is directed in identifying new, less expressed proteins that can deal with important biological functions or conditions.

1.2.2 POST-TRANSLATIONAL MODIFICATIONS (PTMS) ANALYSIS

Almost all proteins are modified from their pure translated amino-acid sequence, by the so-called post-translational modification: there’s a branch of proteomics called protein modification that studies the modified forms of proteins. Post-translational modification (PTM) is the chemical modification of a protein after its translation. It is one of the latest steps in protein biosynthesis for many proteins. A protein (also called a polypeptide) is a chain of amino acids. During protein synthesis, 20 different amino acids can be incorporated in proteins. After translation, the post-translational modification of amino acids extends the range of functions of the protein by attaching to it other biochemical functional groups such as acetate, phosphate, various lipids and carbohydrates, by changing the chemical nature of an amino acid or by making structural changes, like the formation of disulfide bridges. Also, enzymes may remove amino acids from the amino end of the protein, or cut the peptide chain in the middle. Other modifications, like phosphorylation, are part of common mechanisms for controlling the behaviour of a protein, for instance activating or inactivating an enzyme.

Direct analyses of protein modifications are important, since they cannot be predicted from genomic data. Protein modification studies often centre on signal transduction pathways, since signals are most often transmitted by protein modifications such as phosphorylation. There are several types of experiments required for a proteomic approach to study protein modifications. Functional changes of proteins in cells occur because of modification by the attachment of groups such as phosphates, sulphates, carbohydrates, and lipids. There are more than 100 different types of post-translational modifications that can occur to proteins: two of the most important are phosphorylation and glycosylation. Specialized methods have been developed to study phosphorylation (phosphor-proteomics) and glycosylation (glycol-proteomics). Phosphoproteomics is a branch of proteomics that identifies, catalogs, and characterizes proteins containing a phosphate group as a post-translational modification. Glycoproteomics is a branch of proteomics that identifies, catalogs, and characterizes proteins containing carbohydrates as a post-translational modification. Phosphorylation/ glycosylation is a key reversible modification that regulates protein function, sub-cellular localization, complex formation, degradation of proteins and therefore cell signaling networks. With all of these modification results, it is assumed that up to 30% of all proteins may be.

Compared to expression analysis, proteomics provides two additional layers of information. First, it provides clues on what protein or pathway might be activated because a change in phosphorylation/glycosylation status almost always reflects a change in protein activity. Second, it indicates which proteins might be potential drug targets. While proteomics will greatly expand knowledge about the numbers and types of phosphor/glycol-proteins, its greatest promise is the rapid analysis of entire phosphorylation/glycosylation based signaling networks.

1.2.3 PROTEINS TURNOVER

Most of the recent developments in proteomics have focused on improving the technology for protein identification and quantification. Another aspect of protein regulation that must be considered and incorporated into a comprehensive proteomic analysis is protein turnover, the combination of protein synthesis and breakdown. Protein turnover, also known as protein accretion, is the balance between protein synthesis and protein degradation [4]. More synthesis than breakdown indicates an anabolic state that builds lean tissues, whereas more breakdown than synthesis indicates a catabolic state that burns lean tissues. The balance between synthesis and breakdown determines the protein concentration in the cell or tissue. Quantification of proteins in the absence of turnover information may overlook some proteins that are affected by a particular biological condition. For example, the concentration of a protein may not change much, but the rate of turnover can be altered by a condition of interest. In such a situation, the function of the protein may change as older, damaged copies are replaced with newer proteins. A promising approach to solve this problem is to measure the synthesis rate by using in vivo metabolic labeling of proteins with isotope-labeled amino acids and measuring the increment of the protein-bound isotopic enrichment during a study period. The calculation of synthetic rate of the protein also requires the isotopic enrichment in the precursor pool. The technology for large-scale measurement of synthetic rates of individual proteins remains to be established, although some individual protein synthetic rates can be measured in tissue samples. Protein breakdown is also essential to maintain the quality of proteins and their functional integrity. Proteins within cells are continually being degraded to amino acids and replaced by newly synthesized proteins. This is a highly regulated process that prevents accumulation of non-functional and potentially toxic proteins.

A simple and accurate method to study protein breakdown on a protein-by-protein basis has yet to be developed. Protein degradation can be measured across a tissue bed or at whole body level. In vivo measurement of degradation rates of individual proteins is fraught with many problems.

It is important to determine the rate of breakdown of individual protein with a high degree of accuracy and precision to understand the selectivity of the proteolytic process whereby different proteins are committed to breakdown at significantly different rates. Although protein synthesis and breakdown are co-ordinately regulated in the physiological state, their mechanisms are independent. This difference in regulation explains the marked disparity that is sometimes seen between transcriptome and proteome data. For example, changes in mRNA levels can affect protein synthesis, which may or may not result in a change in protein concentration, depending on how protein breakdown is affected.

1.2.4 PROTEIN-PROTEIN INTERACTION

Interaction proteomics concerns the investigation of protein interactions on the atomic, molecular and cellular levels: it is an interesting field because the interaction among proteins is related to all the signalling processes in the cellular regulatory pathways. Thus understanding those proteins interactions networks would help in the comprehension of the molecular signalling. Protein-protein interaction prediction is a field combining bioinformatics and structural biology in an attempt to identify and catalogue interactions between pairs or groups of proteins. Understanding protein-protein interactions is also important in investigating intracellular signalling pathways.

There are many characteristics of a protein-protein interaction that are important. Obviously, it is important to know which proteins are interacting. In many experiments and computational studies, the focus is on interactions between two different proteins. However, you can have one protein interacting with other copies of itself (oligomerization), or three or more different proteins interacting. The stoichiometry of the interaction is also important – that is, how many of each protein involved are present in a given reaction. Some protein interactions are stronger than others, because they bind together more tightly. The strength of binding is known as affinity. Proteins will only bind each other spontaneously if it is energetically favourable. Energy changes during binding are another important aspect of protein interactions. Many of the computational tools that predict interactions are based on the energy of interactions.

Recently there has been a strong focus on predicting protein interactions computationally. Foreseeing the interactions can help scientists to predict pathways in the cell, potential drugs and antibiotics, and protein functions. However, it is a difficult problem. Proteins are large molecules, and binding between them often involves many atoms and a variety of interaction types, including hydrogen bonds, hydrophobic interactions, salt bridges, and more. Proteins are also dynamic, with many of their bonds able to stretch and rotate. Therefore, predicting protein-protein interactions requires a good knowledge of the chemistry and physics involved in the interactions. Consequently protein-protein interaction model are very useful for drug design, since drugs tries to modify, during its clearance time, the biological signalling in order to achieve a therapeutic effect. Bioinformatics and functional proteomic methods take advantage of the known protein structures recorded in the Protein Data Bank database and use information from protein homology, protein functional domains, pathway profiling, and the shape, to model the interaction conformations between two or more proteins and to predict and validate protein complex formation. This approach has been widely used in the computer-aided drug design process. The challenge in this field is the limited number of proteins with known structure because of the difficulties in obtaining enough proteins with crystallographic purity.

1.2.5 BIOMARKERS DISCOVERY

One of the major aim of proteomics is to recognize biomarkers, which are patterns of proteins expression levels that can give a prediction for an early diagnosis, a prognosis or a therapy. The idea is that, since the biological mechanism of life regulation relies on proteic signals, thus, if you understand which will be the system response given a certain proteic signal, then you can predict on the system even if you don’t know at all the complexes regulatory pathways underlying to its working. The biomarker discovery is developed studying the differential proteic expression comparing, for instance, healthy vs. unhealthy subjects. It borrowed basically the methods implemented for the gene differential expression analysis. For some poor prognostic malignancies, such as pancreatic and ovarian cancers, early diagnosis and surgery are the best therapeutic approaches. There are no specific and highly sensitive biomarkers available for these diseases. A self-trained pattern recognition algorithm has been proven capable of identifying proteomic patterns in MS signal to completely segregate cancer from normal, although no specific proteins were identified. These pattern recognition algorithms involve complicated neural networking technologies, but it is needed a specificity and sensitivity increase. The specificity is a statistical measure of how well a binary classification test correctly identifies the negative cases, or those cases that do not meet the condition under study. For example, given a medical test that determines if a person has a certain disease, the specificity of the test to the disease is the probability that the test indicates "negative" if the person does not have the disease. That is, the specificity is the proportion of true negatives to all negative cases in the population. It is a parameter of the test. High specificity is important when the treatment or diagnosis is mentally and/or physically harmful for the patient. Sensitivity, or recall rate, is a statistical measure of how well a binary classification test correctly identifies a condition, whether this is medical screening tests picking up on a disease. The results of the screening test are compared to some absolute gold standard; for example, for a medical test to determine if a person has a certain disease, the sensitivity to the disease is the probability that if the person has the disease, the test will be positive. The sensitivity is the proportion of true positives of all diseased cases in the population. It is a parameter of the test. High sensitivity is required when early diagnosis and treatment is beneficial, and when the disease is infectious.

Several techniques allow to test for proteins produced during a particular disease, which helps to diagnose the disease quickly. Techniques include western blot, immunohistochemical staining, enzyme linked immunosorbent assay (ELISA) or mass spectrometry. If proteomics will detect a set of biomarkers for every disease it will be easier, more comfortable and time-earning to make a diagnosis, in the brightest occurrence it will be possible just analyzing the serum. Thus it could be also money saving for the hospitals and a real pre-diagnosis will be likely for all the population. The research is still working on it, but interesting results have been reached until now. Most studies deal with cancer: ovarian cancer, prostate cancer, breast cancer, kidney cancer, colon cancer. For instance, proteomic analysis of kidney cells and cancerous kidney cells is producing promising leads for biomarkers for renal cell carcinoma and developing assays to test for this disease. In kidney-related diseases, urine is a potential source for such biomarkers. Recently, it has been shown that the identification of urinary polypeptides as biomarkers of kidney-related diseases allows to diagnose the severity of the disease several months before the appearance of the pathology.

In Alzheimer’s disease, elevations in beta secretase creates amyloid/beta-protein, which causes plaque to build up in the patient's brain, which causes dementia. Targeting this enzyme decreases the amyloid/beta-protein and so slows the progression of the disease. A procedure to test for the increase in amyloid/beta-protein is immunohistochemical staining, in which antibodies bind to specific antigens or biological tissue of amyloid/beta-protein.

Heart disease is commonly assessed using several key protein based biomarkers. Standard protein biomarkers for CVD include interleukin-6, interleukin-8, serum amyloid A protein, fibrinogen, and troponins. cTnI cardiac troponin I increases in concentration within 3 to 12 hours of initial cardiac injury and can be found elevated days after an acute myocardial infarction. A number of commercial antibody based assays as well as other methods are used in hospitals as primary tests for acute MI. We hope in the future to develop similar proteomic based tests for all the diseases.

1.2.6 PROTEOMICS FOR DRUGS DISCOVERY

The recent boom of the proteomics field, or the analysis of the ever dynamic proteome, has brought many advances with respect to the very nature of how the current drug discovery process is undertaken. The potential the field of proteomics brings in, for identifying proteins involved in disease pathogenesis and physiological pathway reconstruction, facilitates the ever increasing discovery of novel drug targets, their respective modes of action mechanistically, and their biological toxicology.

The challenge in the drug discovery process is to find the exact causes of an underlying disease and find a way to negate them or bring them to normal levels. A mechanistic understanding of the nature of the disease in question is essential if we aim at elucidating any target-specific remedy for it. While the causes of many documented clinical problems greatly vary in their nature and origin, the consequences are mostly found at the protein level, involving protein function, protein regulation, or protein-protein interactions. Indeed, identification of potential new drugs for the treatment of disease relies on genome and proteome information to identify proteins associated with a disease. For example, if a certain protein is implicated in a disease, its 3D structure provides the information to design drugs to interfere with the action of the protein. A molecule that fits the active site of an enzyme, but cannot be released by the enzyme, will inactivate the enzyme. This is the basis of new drug-discovery tools, which aim to find new drugs to inactivate proteins involved in disease.

Recent advances in applied genomics helped in the target identification process, since it allowed for high throughput screening of expressed genes. As genetic differences among individuals are found, researchers expect to use these techniques to develop personalized drugs that are more effective for the individual. However, studies have shown that there is a poor correlation between the regulation of transcripts and actual protein quantities. The reasons for this are that genome analysis couldn't account for post-translational processes such as protein modifications and protein degradation. Therefore, the methods employed in the drug-discovery process started to shift from genomics to proteomics. Analysis of the dynamic proteome, as opposed to the static genome, will certainly bring a much more accurate approach to identify not only applicable biomarkers that will aid in diagnosis, but also effective remedies for diseases of varying origins.

The field of proteomics faces some daunting challenges, in comparison to genomics, for several reasons. First, protein science lacks an analogue of the polymerase chain reaction (PCR), which can generate many copies of a single, native molecule in vivo (nucleic acids in the case of PCR). However, several recent approaches have been applied in an effort to ameliorate the situation. Methods of chemical synthesis exist, being limited by yield, particularly when it comes to synthesizing lengthy peptides. In-vivo expression synthesis methods exist as well, however, this approach cannot be applied to producing proteins which may alter normal cellular function. Also, cell-free synthesis ribosome kits can be employed for accurate and rapid protein synthesis, though the intrinsic presence of ribosome inactivating enzymes contributes to the instability of these systems. Second, in contrast to DNA, protein levels vary significantly depending on cell type and environment. Third, protein abundance is not directly correlated to protein activity, which, in fact, is often determined by post-transcriptional modifications such as phosphorylation

The ideal proteomics technique suited for drug discovery would have the following features: it should be able to separate membrane proteins and detect low abundance proteins, two abilities not quite yet realized, yet required in current separations and analytical techniques. Furthermore, it should be able to identify protein activity independently of protein abundance. It also should reveal protein-protein and protein-small-molecule interactions. This method should also be implemented easily, be automatable, and perform at high-throughput speed. Proteomics researchers are addressing these issues, and new methods are being developed.

1.2.7 CELLULAR PROTEOMICS AND SYSTEMS BIOLOGY

Cellular proteomics is a new branch of proteomics aiming to map the location of proteins and protein-protein interactions in whole cells during key cell events. It uses techniques such as X-ray Tomography and optical fluorescence microscopy.

Systems biology has been enabled by recent advances in multi-disciplinary scientific disciplines that allow for the parallel large-scale measurement of biomolecules, such as mRNA, proteins and metabolites. Understanding the detailed physiology of cells, tissues and entire organisms afforded by this approach will lead to a more comprehensive understanding of basic cellular events and their coordination. This comprehensive investigative approach represents a major shift in scientific paradigm, and over time will clearly have a major impact on how scientific analysis will be conducted.

The recent few years have seen a growing interest in defining and establishing the emerging discipline of systems biology. While it is difficult to clearly define such a rapidly evolving discipline, characteristic trends are becoming apparent that allow a definition of what systems biology plans to accomplish. System biology endeavours to understand the detailed coordinated workings of entire organisms, with the ultimate goal to detect differences between health and disease, or to understand how cells or entire organisms react to the environment. Its ultimate goal is to understand the dynamic networks of regulation and interactions that allows cells and organisms to live in a highly interactive environment, and to understand how perturbations in the system cause disease.

The critics of systems biology are ready to point out that "omic" approaches are not a substitute for hypothesis driven research, because a systems analysis does not provide a testable hypothesis but is more like a "fishing expedition", yielding undetermined information of a collective of molecules. However, this view-point does not do the discipline justice, because large scale investigative approaches can be hypothesis driven. For instance, one can form more global hypotheses such as a cell line or tissue changes protein expression/modification patterns in response to a drug stimulus, and that these changes are causally related to a toxic response to the drug. Using integrated molecular tools, these induced changes can readily be measured and compared to an appropriate experimental control. Cluster and correlation analysis of these data will then readily describe the dynamics of molecular changes in response to a perturbation of the system, in this case a drug challenge. Taken at face value, this collective information will provide the researcher with a foundation to create better-informed hypotheses. This then accelerates the discovery process by avoiding the sequential trial and error approach that often plague classical experimentation. The real issue is that high-throughput approaches, such as gene expression analysis, proteomics, and metabolomics (the quantification and identification of metabolites and their modifications) provide only part of the cellular picture, namely the collective of molecules in a cell.

While the integration of all omics information can provide great insights into how genetic and proteomic programs are modulated, the information alone does not provide any mechanistic details of how these molecules catalyze chemical reactions. The latter information can only be obtained through reductionist approaches, for example through the structural and functional analyses of proteins and the reconstitution of biological processes in vitro, which can scientifically prove function and mechanism. Knowledge about tissue specific and subcellular protein localization, together with quantitative information about local or cellular abundance, will add further detail that allows the interpretation and assessment of which machineries are localized where and if a given mechanism is likely to be significant to a particular process.

2 MASS SPECTROMETRY-BASED PROTEOMICS

Mass spectrometry (MS)-based proteomics, providing information about the qualitative and quantitative content of a biological sample, has become the technique of choice for acquiring data in the proteomics research field. This chapter is meant to illustrate all the main steps of a MS-based proteomics workflow.

2.1 SAMPLE PREPARATION

As technological progresses are made in the field of proteomics, it is seen that advances are necessary in the preparation of protein samples. Over time, changes will take place in protein structure that could potentially alter experimental results; many problems can arise from improper handling of proteins. Contaminants in samples can cause results to be skewed, and may even damage equipment. Uneven labeling could compromise quantification reliability.

Sample preparation is becoming particularly critical in the case of high throughput techniques involving Mass spectrometry (MS). In these protocols, the conditions of a sample in one stage may directly conflict with the efficacy of a second stage. A number of issues arise in this respect; including sample fractionation, clean-up, labeling, etc. Thus, there is an increasing demand for automated and streamlined sample preparation tools for protein and peptide extraction upstream of MS. The particular MS experiment will ultimately dictate the degree and types of necessary preparations. Nevertheless, some concepts, such as fractionation, can be applied to any MS experiment, whereas labeling is used only for relative quantification, which will be illustrated in the next chapter.

Therefore, the first step to any proteomics experiment, particularly MS, is to reduce the complexity of the sample, or fractionate the sample. The goal of fractionation is usually to remove the “highly abundant components of the proteome” followed by “subsequent fractionation of the moderate to low abundance proteins” in order to produce a concentrated sample of proteins with the potential to be clinically relevant. A protein sample can be fractionated on the basis of size, charge, hydrophobicity, and/or binding affinity. These qualities are often the basis of the many available kits capable of enriching a sample by partitioning out highly abundant proteins from a sample such as serum.

The clean-up of a protein sample is more crucial prior to introduction to the mass spectrometer than in most other proteomics experiments, primarily because of the sensitivity of the technology, but also because dirty samples can be quite detrimental to the machine. In terms of MS sensitivity, it is necessary to remove the detergents, ion suppressing salts, and other substances commonly used in proteomic sample preparation as they can compromise analysis.

Protein samples need to be denatured prior to any digestion with proteases so the protease will have as much access to targeted amino acids as possible [5]. Proteases cleavage is needed for reducing protein to peptides, which can be detected by the MS since the m/z ratios of their isotopes belong to the mass acquisition range of the spectrometer. Endoproteinase Lys-C (Lys-C) and trypsin are proteases used for digesting proteins into a population of peptides that can be identified by the mass spectrometer. Lys-C cleaves on the c-terminal side of lysine and the resulting peptides are larger than tryptic peptides. Trypsin has a high specificity, it cleaves on the c-terminal side of lysine and arginine amino acids. Since maximal amino acid coverage of the protein is required, it is best to digest the sample with several proteases, so the resulting peptides are more likely to contain amino acid information from the entirety of the protein. The selection of proteases depends greatly upon the amino acid sequence of the target protein.

2.2 PROTEIN SEPARATION

In chemistry and chemical engineering, a separation process is used to transform a mixture of substances into two or more compositionally-distinct products. Almost every element or compound is found naturally in an impure state such as a mixture of two or more substances. To obtain a pure protein sample, a protein must be isolated from all other proteins and cellular components. A task that is equally challenging is keeping the protein in its active form. When purifying proteins it is necessary to simulate the pH, salt concentration and reducing conditions in which they normally are. In the process of obtaining an active and pure sample it is convenient to minimize the number of steps taken in order to maximize the yield at the end of the separation. Finally, since proteins are subject to fast degradation rates, it is also critical to obtain our sample as quickly as possible. All these components of protein separations can be successfully achieved by a group of separation methods collectively known as chromatography. There are other separation techniques, e.g., electrophoresis and centrifugation, but chromatography is of utmost importance for MS-based proteomics research. In next paragraphs, some of the chromatographic techniques commonly coupled to MS will be described.

2.2.1 CHROMATOGRAPHIC SEPARATION

Chromatography makes use of an insoluble stationary phase and a mobile phase: the mobile phase is commonly a liquid solution which contains the protein to be isolated. The stationary phase on the other hand is made up of a grouping of beads, usually based on a carbohydrate or acrylamide derivative, that are bound to ionic charged species, hydrophobic characters, or affinity ligands. Successful chromatography design depends upon the selection of an appropriate stationary phase. There are several properties of proteins that can be taken advantage of for separating proteins. Different types of chromatography take advantage of different properties. Proteins can be separated by size, shape, hydrophobicity, affinity to molecules or charge. The most common form of chromatography used in proteomics is probably Column chromatography. In column chromatography (represented in Figure 2-1), a mixture of proteins in solution is applied to the top of a cylindrical column filled with a permeable solid matrix immersed in solvent. A large amount of solvent is then pumped through the column. Depending on the type of chromatography, proteins with certain characteristics will bind to the stationary phase while those lacking the sought characteristics will remain in the mobile phase and pass through the column. The final step involves displacing the protein from the stationary phase, also known as elution, by introducing a particle which will compete with the protein binding site on the stationary phase. Because different proteins are retarded to different extents by their interaction with the matrix, they can be collected separately as they flow out from the bottom. Indeed, the column is usually coupled to a detection device such as a mass spectrometry device. Today various commercial column are readily available. The mobile phase can be either liquid or gas.

Gas chromatography is very widely used in analytical chemistry. It has less application to proteomics because the technique requires high temperatures which are often unsuitable for the large polymers involved in proteomics. Gas chromatography depends on the partition equilibrium between a solid stationary phase and a gaseous mobile phase. It is almost always performed in a tube. The stationary phase usually consists of solid beads packed into a column adhered to a capillary tube.

A more useful mobile phase in proteomics is one that is in a liquid state. All of the techniques discussed in this paragraph involve liquid chromatography. In this technique, there is traditionally a partition equilibrium between a solid stationary phase and a liquid mobile phase. Liquid chromatography is either carried out in a column or a plane. The stationary phase is almost always solid, however, there are examples of chromatography experiments in which the stationary phase is in another state.

Separation of highly complex mixture can be a very difficult task. The mixture can be distributed according to their molecular mass, chemical composition, functionality and architecture. A single chromatography experiment may be inefficient in separating our proteins of interest. In 2D chromatography, different techniques are essentially combined to achieve a higher degree of separation. This can be done by an offline technique, where the result of one chromatography is injected manually into a second column chromatography or an online method, where the two columns are directly coupled through switches.

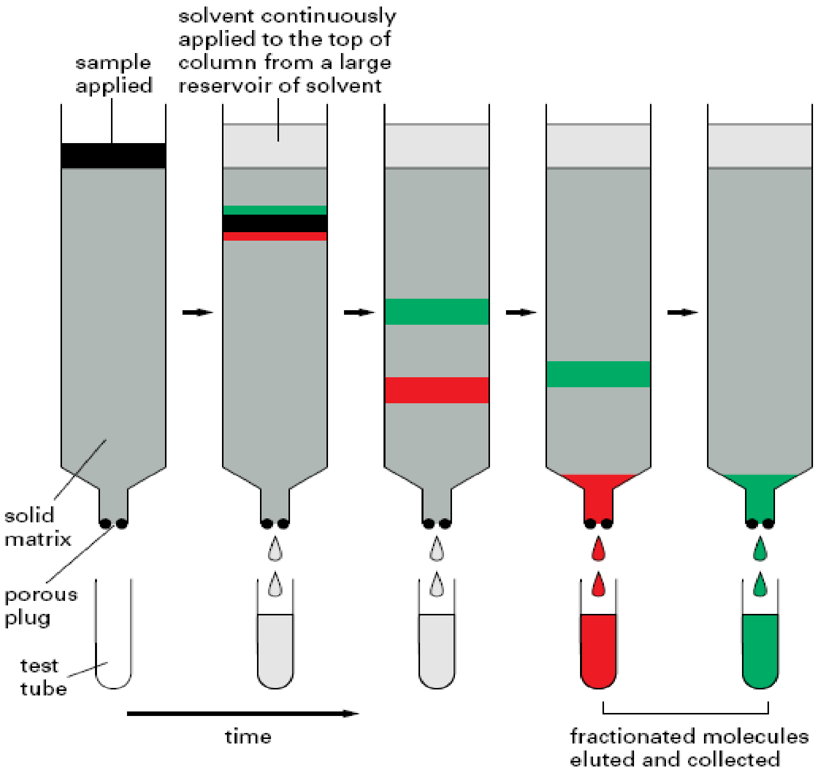

solvent continuously applied to the top of column from a large reservoir of solvent
sample applied
solid matrix
porous plug
test tube
time
fractionated molecules eluted and collected

FIGURE 2-1 A CHROMATOGRAPHIC COLUMN SYSTEM WITH SOLID MATRIX. THE STATIONARY PHASE IS IN A COLUMN. THE MOBILE PHASE ENTERS THE COLUMN AND FLOWS OUT AT A CONSTANT RATE. AS IT FLOWS OUT OF THE COLUMN ANY PROTEIN THAT HASBEEN ELUTED IN THE MOBILE PHASE CAN BE DETECTED.

2.2.1.1 ION EXCHANGE CHROMATOGRAPHY

Ion exchange chromatography (IC) is probably the most frequently used chromatographic technique for the separation and purification of proteins, polypeptides, nucleic acids, polynucleotides and other charged biomolecules based on the charge properties of the molecules. The reasons for the success of ion exchange are its widespread applicability, its high resolving power, its high capacity and the simplicity and controllability of the method. Ion exchange chromatography retains analyte molecules based on ionic interactions (see Figure 2-2). The stationary phase surface displays ionic functional groups that interact with analyte ions of opposite charge. The charged stationary phases are named according to the types of charged particles that bind to them. This type of chromatography is further subdivided into cation exchange chromatography and anion exchange chromatography:

- Cation-exchange chromatography retains positively charged cations because the stationary phase displays a negatively charged functional group such as a phosphoric acid;
- Anion-exchange chromatography retains negatively charged anions using positively charged functional group such as a quaternary ammonium cation.

Proteins have numerous functional groups that can have both positive and negative charges. Ion exchange chromatography separates proteins according to their net charge, which is dependent on the composition of the mobile phase. By adjusting the pH or the ionic concentration of the mobile phase, various protein molecules can be separated. For example, if a protein has a net positive charge at pH 7, then it will bind to a column of negatively-charged beads, whereas a negatively charged protein would not. By changing the pH so that the net charge on the protein is negative, it will be eluted too.

Elution by changing the ionic strength of the mobile phase is a more subtle effect: it works as ion from the mobile phase will interact with the immobilized ion in preference over those on the stationary phase. This shields the stationary phase from the protein binding (and vice versa) and allows the protein to elute.

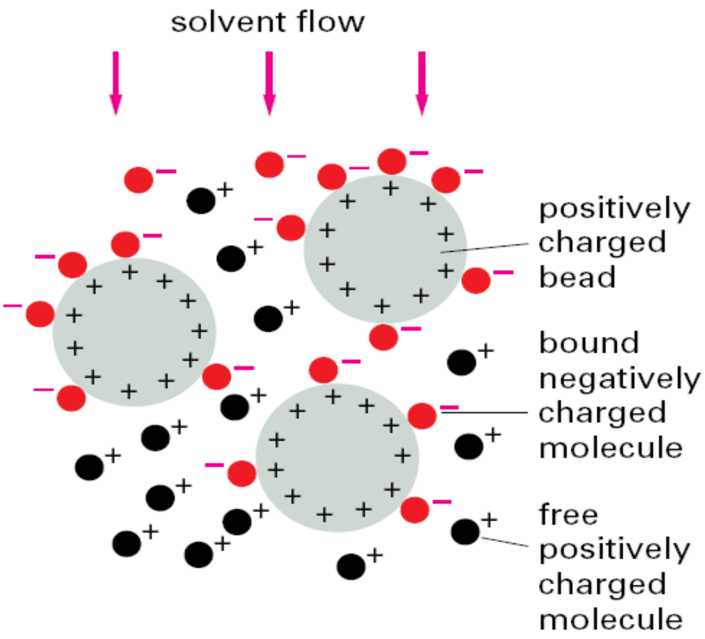

solvent flow
positively charged bead
bound negatively charged molecule
free positively charged molecule

FIGURE 2-2 ION-EXCHANGE COLUMNS ARE PACKED WITH SMALL BEADS CARRYING EITHER POSITIVE OR NEGATIVE CHARGES THAT RETARD PROTEINS OF THE OPPOSITE CHARGE. THE ASSOCIATION BETWEEN A PROTEIN AND THE MATRIX DEPENDS ON THE PH AND IONIC STRENGTH OF THE SOLUTION PASSING DOWN THE COLUMN.

2.2.1.2 AFFINITY CHROMATOGRAPHY

Affinity chromatography is one of the most commonly used techniques as it is very selective and effective at isolating proteins. The technique relies on unique interaction between a molecules with a ligand bounded to the matrix (see Figure 2-3). These matrices include interaction between those pairs: antigen-antibody, enzyme-substrate, receptor-ligando, nucleic acid binding protein-nucleic acid and polysaccharide/glycoprotein-lectin. Developing an effective affinity chromatography method involves finding a ligand that is specific enough and creating suitable conditions for the binding between the target protein and the ligand as well as to release the protein. Since only the specific target sample can bind to the stationary phase, no fine-tuned elution gradient is necessary.

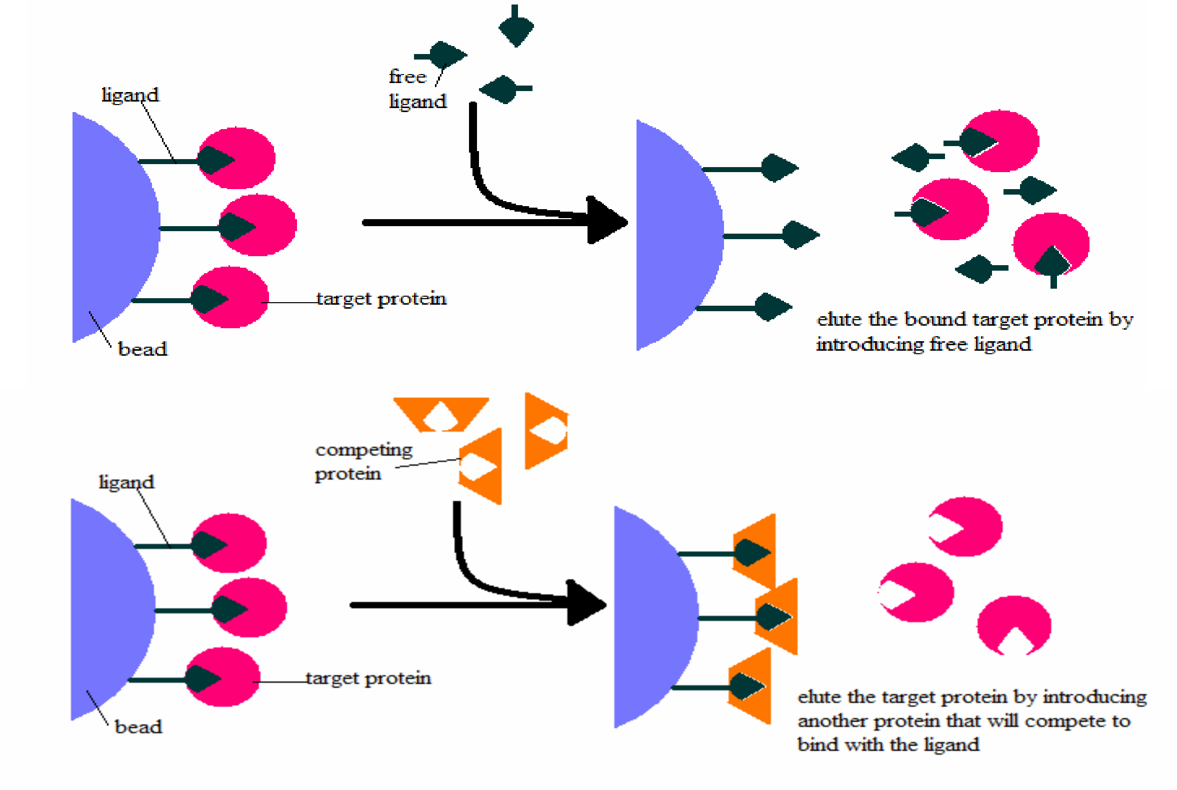

ligand
free ligand
target protein
bead
elute the bound target protein by introducing free ligand
ligand
competing protein
target protein
bead
elute the target protein by introducing another protein that will compete to bind with the ligand

FIGURE 2-3 ONE WAY TO MAKE THE BOUND PROTEIN ELUTE IS TO INTRODUCE FREE LIGAND THAT WILL BIND TO THE TARGET MOLECULE (UPPER PANEL). THE BOUND PROTEIN CAN BE ELUTED BY INTRODUCING ANOTHER PROTEIN THAT WILL OUTCOMPETE THE TARGET PROTEIN AND BIND TO THE LIGAND (PANEL BELOW).

2.2.1.3 NORMAL PHASE CHROMATOGRAPHY

Normal phase chromatography (NP) separates analytes based on polarity. This method uses a polar stationary phase and a non-polar mobile phase, and is used when the analyte of interest is fairly polar in nature. The polar analyte associates with and is retained by the polar stationary phase. Adsorption strengths increase with increase in analyte polarity, and the interaction between the polar analyte and the polar stationary phase (relative to the mobile phase) increases the elution time. Use of more polar solvents in the mobile phase will decrease the retention time of the analytes while more hydrophobic solvents tend to increase retention times. Particularly polar solvents in a mixture tend to deactivate the column by occupying the stationary phase surface. This is somewhat particular to normal phase because it is most purely an adsorptive mechanism (the interactions are with a hard surface rather than a soft layer on a surface).

NP chromatography had fallen out of favour in the 1970's with the development of reversed-phase chromatography because of its lack of reproducibility of retention times.

2.2.1.4 REVERSED PHASE CHROMATOGRAPHY

Reversed Phase chromatography (RP) is a separation technique based on the solubility of the protein. The term “reverse” was derived from its predecessor named “normal” phase chromatography, which utilized a polar stationary phase. In reverse phase, the stationary phase is packed with non-polar hydrocarbon, typically C4, C8 or C18. This creates a hydrophobic stationary phase, in contrast with the polar stationary phase of the NP. The mobile phase on the other hand, contains polar organic solvents such as methanol, butanol, isopropanol, acetonitrile and isopropanol. Utilization of these polar solvents introduces very harsh conditions for the protein, thus the method will generally work well for smaller and more stable proteins. All peptides and proteins carry a mix of hydrophilic and hydrophobic amino acids, but those with high net hydrophobicity will be able to participate in hydrophobic interactions with the stationary phase. As mixtures of proteins are applied to the column, polar proteins will elute first while non-polar proteins will bind to the column. Proteins in the mixture that have a high percentage of exposed hydrophobic amino acid residues will be adsorbed to the hydrophobic stationary phase. Other proteins in the mixture will be washed out. Elution of the bound hydrophobic protein can be accomplished by increasing the concentration of organic solvent, which increases the retention time of a particular component. Reverse phase chromatography is commonly coupled with mass spectrometry in an effort to quantify the protein that is eluted from the column and is the method used to generate the dataset we will analyze in this work. For the sake of exhaustiveness the dataset have been separated using nanoRP-HPLC applied on a Ultimate 2 Dual Gradient HPLC system.

2.2.1.5 HIGH-PERFORMANCE/PRESSURE LIQUID CHROMATOGRAPHY

High-performance/pressure liquid chromatography (HPLC) is a form of column chromatography used frequently in biochemistry and analytical chemistry. HPLC is used to separate components of a mixture by using a variety of chemical interactions between the substance being analyzed (analyte) and the chromatography column. The basic operating principle of HPLC is to force the analyte through a column of the stationary phase (usually a tube packed with small round particles with a certain surface chemistry) by pumping a liquid (mobile phase) at high pressure through the column. The internal diameter (ID) of an HPLC column is a critical aspect that determines quantity of analyte that can be loaded onto the column and also influences sensitivity. Larger columns are usually seen in industrial applications, low ID columns have improved sensitivity and lower solvent consumption at the expense of loading capacity. The sample to be analyzed is introduced in small volume to the stream of mobile phase and is retarded by specific chemical or physical interactions with the stationary phase as it traverses the length of the column. The amount of retardation depends on the nature of the analyte, stationary phase and mobile phase composition. The time at which a specific analyte elutes (comes out of the end of the column) is called the retention time and is considered a reasonably unique identifying characteristic of a given analyte. The use of pressure increases the linear velocity (speed) giving the components less time to diffuse within the column, leading to improved resolution in the resulting chromatogram (that is the temporal representation of the eluting substance). Common solvents used include any miscible combinations of water or various organic liquids (the most common are methanol and acetonitrile). Water may contain buffers or salts to assist in the separation of the analyte components, or compounds such as Trifluoroacetic acid which acts as an ion pairing agent.

A further refinement to HPLC has been to vary the mobile phase composition during the analysis, this is known as gradient elution. A normal gradient for reversed phase chromatography might start at 5% methanol and progress linearly to 50% methanol over 25 minutes, depending on how hydrophobic the analyte is. The gradient separates the analyte mixtures as a function of the affinity of the analyte for the current mobile phase composition relative to the stationary phase. This partitioning process is similar to that which occurs during a liquid-liquid extraction but is continuous, not step-wise. In this example, using a water/methanol gradient, the more hydrophobic components will elute (come off the column) under conditions of relatively high methanol; whereas the more hydrophilic compounds will elute under conditions of relatively low methanol. The choice of solvents, additives and gradient depend on the nature of the stationary phase and the analyte. Often a series of tests are performed on the analyte and a number of generic runs may be processed in order to find the optimum HPLC method for the analyte, which gives the best separation of peaks. Most traditional HPLC is performed with the stationary phase attached to the outside of small spherical silica particles (very small beads). These particles come in a variety of sizes with 5µm beads being the most common. Smaller particles generally provide more surface area and better separations, but the pressure required for optimum linear velocity increases by the inverse of the particle diameter squared. This means that changing to particles that are half as big in the same size of column will double the performance, but increase the required pressure by a factor of four. High performance liquid chromatography has proven itself to be very useful in many scientific fields, yet forces scientists to consistently choose between speed and resolution.

2.2.1.6 ULTRA PERFORMANCE LIQUID CHROMATOGRAPHY

Ultra performance liquid chromatography (UPLC or uHPLC) eliminates the need to choose and creates a highly efficient method that is primarily based on small particle separations. uHPLC systems have been developed to take into account all the advantages that small particle separations currently have over HPLC. Many of these advantages are primarily based on the theories behind liquid chromatography. In general, increasing the efficiency of a separation will also increase its resolution. Since both efficiency and optimum flow rate are inversely proportional to particle size, a decrease in the particle size will increase efficiency and speed up the flow rate. The particles are specifically designed to withstand wide ranges of pressure and pH, have a high load capacity, and improve efficiency. Other innovations to the chromatography method include a high pressure solvent delivery system, to take into account the smaller particle size, fast injection cycle sample management, and specialized detectors with fiber optic flow cell design. The lower bead size is the true reason for uHPLC increased flow rate and resolution. This can be shown mathematically using Deemter's equation: H = A + B/μ + Cμ. H being the plate height and μ being the particle size. The A, being a constant, is independent of flow rate (it is referred to as the "Eddy diffusion term"). The B constant is the diffusion coefficient, and C is the "analyte mass transfer" coefficient. As μ decreases, the A and C values needed for a similar H value decrease, allowing for higher resolution. This also reduces the effect of the C value on the H value, yielding faster separations for similar resolutions. Note uHPLC out classes HPLC in all aspects, and is expected to replace HPLC in the near future.

2.2.2 SEPARATION BY ELECTROPHORESIS

Gel electrophoresis is used to differentiate molecular entities depending on their physical characteristics such as size, shape, or isoelectric point as they move through a gel by an electrical current. Gel electrophoresis is used as an analytical technique or as a preparatory technique to purify molecules before they are used for other methods like mass spectrometry. It is based on the principle that, when charged molecules are placed in an electric field, they migrate toward either the positive or negative pole depending on their charge. Since nucleic acids are negatively charged due to their phosphate groups they migrate toward the anode. Unlike nucleic acids, since proteins can have either a net positive or a net negative charge they can migrate to either of the poles depending on the charge. Protein can have different charges and complex shapes, primary, secondary, tertiary, and quaternary structure and that make migration through the gel have extremely different rates during electrophoresis.

2.2.2.1 POLYACRYLAMIDE GEL ELECTROPHORESIS

Polyacrylamide gel electrophoresis (PAGE) is commonly used separating proteins. PAGE can be used to purify proteins prior to other proteomics techniques or to analyze information on the mass, the charge on proteins, and/or presence of a protein. Due to these complex structures, proteins are usually denatured, or broken down to simple primary structures in the presence of a detergent such as sodium dodecyl sulfate (SDS), which imparts a negative charge on proteins, and thus allow for proper migration. The quantity of SDS bound and the size of the protein are relative to each other, thus this method separates proteins mainly based on molecular weight.

Two-dimensional PAGE (2-D PAGE) differentiates proteins in the first dimension by isoelectric point and in the second dimension by molecular weight. Native PAGE separates proteins by mass/charge ratio without denaturing them.

2.2.2.2 SDS-PAGE

SDS-PAGE is a very common method of gel electrophoresis for separating proteins by mass. It was first employed by U.K Laemmli and known as Laemmli method. The proteins are dissolved in sodium dodecyl sulfate (SDS), a detergent that breaks up the interactions between proteins, and then electrophorised. The smallest molecules move through the gel faster, while larger molecules take longer and result in bands closer to the top of the gel. The gel used for SDS-PAGE is made out of acrylamide, which forms cross-linked polymers of polyacrylamide. Standard gels are typically composed of two layers, the stacking gel (top layer) and separating or resolving gel (lower layer). The stacking layer contains a low percentage of acrylamide and has low pH, while the acrylamide concentration of the separating gel varies according to the samples to be run and has higher pH. The differences in pH and acrylamide concentration at the stacking and separating gel provide better resolution and sharper bands in the separating gel.

The gel is submerged in the buffer and proteins denatured by SDS are applied to one end of a layer of gel. Buffer provides uniform pH and ions for conducting electric potential. The proteins which are negatively charged migrate across the gel to the positive pole when an electricity is applied through the gel. Short proteins move fast because they can easily pass through the gel pores, while larger molecules move slowly. Due to differential migration based on their size, larger proteins are close to the top of the gel while smaller proteins move to bottom of the gel. After a given period of time, proteins might have separated roughly according to their sizes. Proteins of known molecular weight (marker proteins) can be run in a separate lane in the gel for calibration.

After the electrophoresis run, the gel is stained with silver stain or Coomassie Brilliant Blue for visualization of the proteins. Within the gel different proteins will be seen as separate spots or bands depending on their sizes on staining. The molecular weight of a protein in the band can be estimated by comparing it with the marker proteins of known molecular weights. The separated proteins can be cut from the gel and further analyzed by other proteomics techniques.

2.2.2.3 2D ELECTROPHORESIS

2-D electrophoresis begins with 1-D electrophoresis but then separates the molecules by a second property in a direction 90 degrees from the first. In 1-D electrophoresis, proteins (or other molecules) are separated in one dimension, so that all the proteins/molecules will lie along a lane but be separated from each other by a property (e.g. isoelectric point). The result is that the molecules are spread out across a 2-D gel. Because it is unlikely that two molecules will be similar in both properties, molecules are more effectively separated in 2-D electrophoresis than in 1-D electrophoresis. However 1-D gel electrophoresis (e.g. SDS-PAGE) is more commonly used. The two dimensions that proteins are separated into using this technique are isoelectric point and mass. To separate the proteins by isoelectric point, a gradient of pH is applied to a gel and an electric potential is applied across the gel, making one end more positive than the other. At all pHs other than their isoelectric point, proteins will be charged. If they are positively charged, they will be pulled towards the more negative end of the gel and if they are negatively charged they will be pulled to the more positive end of the gel. The proteins applied in the first dimension will move along the gel and will accumulate at their isoelectric point. That is, the point at which the overall charge on the protein is 0 (i.e. a neutral charge). Before separating the proteins by mass, they are treated with sodium dodecyl sulfate along with other reagents (SDS-PAGE in 1-D). This denatures the proteins (that is, it unfolds them into long, straight molecules) and binds a number of SDS molecules roughly proportional to the protein's length. Because a protein's length (when unfolded) is roughly proportional to its mass, this is equivalent to saying that it attaches a number of SDS molecules roughly proportional to the protein's mass. Since the SDS molecules are negatively charged, the result of this is that all of the proteins will have approximately the same mass-to-charge ratio as each other. In addition, proteins will not migrate when they have no charge (a result of the isoelectric focusing step) therefore the coating of the protein in SDS (negatively charged) allows migration of the proteins in the second dimension (SDS is not compatible for use in the first dimension as it is charged and a nonionic or zwitterionic detergent needs to be used). In the second dimension, an electric potential is again applied, but at a 90 degree angle from the first field. The proteins will be attracted to the more positive side of the gel proportionally to their mass-to-charge ratio. As previously explained, this ratio will be nearly the same for all proteins. The migration will be slowed by frictional forces. The gel therefore acts like a molecular filter when the current is applied, separating the proteins on the basis of their molecular weight with larger proteins being retained higher in the gel and smaller proteins being able to pass through the sieve and reach lower regions of the gel. The result is a gel with proteins spread out on its surface. These proteins can then be detected by a variety of means, but the most commonly used stains are silver and Coomassie staining (see Figure 2-4). In this case, a silver colloid is applied to the gel. The silver binds to cysteine groups within the protein. The silver is darkened by exposure to ultra-violet light. The darkness of the silver can be related to the amount of silver and therefore the amount of protein at a given location on the gel. This measurement can only give approximate amounts, but is adequate for most purposes.

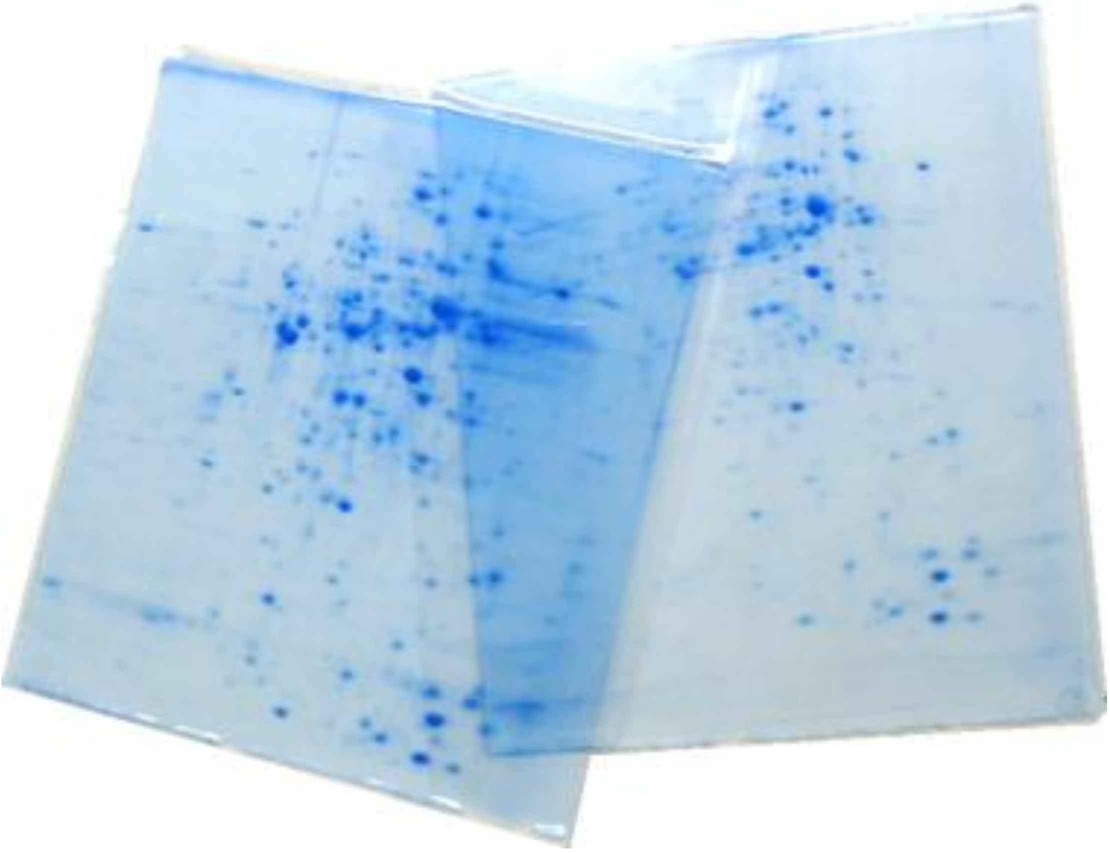


FIGURE 2-4 COOMASSIE STAINED 2D GELS FOR 2D ELECTROPHORESIS.

2.2.3 SEPARATION BY CENTRIFUGATION

Centrifugation is one of the most important and widely applied research techniques in biochemistry, cellular and molecular biology, and in medicine. In the field of proteomics it plays a vital role in the fundamental and necessary process of isolating proteins. This process begins with intact cells or tissues. Before the proteins can be obtained, the cells must be broken open by processes such as snap freezing, homogenization by high pressure, or grinding with liquid nitrogen. Once the cells have been opened up all of their contents; including cell membranes, RNA, DNA, and organelles will be mixed in the solvent with the proteins. Centrifugation is probably the most commonly used method for separating out all the non proteic material. Within the centrifuge samples are spun at high speeds and the resulting force causes particles to separate based on their density. Moreover the use of density gradients externally applied has become almost routine in centrifugal fractionation of particle mixtures and purification of subcellular organelles and macromolecules. The basic idea behind the density gradient approach is that the mixture of particles to be separated is placed onto the surface of a vertical column of liquid, the density of which progressively increases from top to bottom, and then centrifuged. Although the particles in suspension are individually denser than the liquid at the top of the gradient, the average density for the sample (i.e., particles plus suspending liquid) is lower; only under such conditions could the sample zone be supported by the top of the density gradient. We won't go through this kind of separation techniques anymore since our focus is chromatography. For more detailed information about it see some of the countless review on it.

2.3 MASS SPECTROMETRY

Mass spectrometry (MS) is an analytical technique used to measure the mass-to-charge ratio of ions. The technique had its beginnings in J.J. Thomson's vacuum tube where, in the early part of the century, the existence of electrons and "positive rays" was demonstrated. Thomson, the physicist, observed in his book "Rays of Positive Electricity and Their Application to Chemical Analysis" that the new technique could be used profitably by chemists to analyze chemicals. Despite this far-sighted observation, the primary application of mass spectrometry remains in the realm of physics for nearly thirty years. It was used to discover isotopes, to determine their relative abundance, and to measure their exact atomic masses, with a precision of 1 part in 10^6 or better. These important fundamental measurements laid the foundation for later developments in different fields ranging from geochronology to biochemical research.

MS is used to find the composition of a physical sample by generating a mass spectrum representing the relative concentrations (i.e. intensities) of the masses of sample components (see Figure 2-5). The mass spectrum is measured by the mass spectrometer. More specifically, a mass spectrometer is an instrument that measures the masses of individual molecules that have been converted into ions, i.e., molecules that have been electrically charged. The unit of mass is often referred to by chemists and biochemists as the Dalton (Da for short), and is defined as follows: 1 Da=(1/12) of the mass of a single atom of the isotope of carbon-12 (^{12}C). This follows the accepted convention of defining the ^{12}C isotope as having exactly 12 mass units. A mass spectrometer does not actually measure the molecular mass directly, but rather the mass-to-charge ratio of the ions formed from the molecules. A useful unit for this purpose is the fundamental unit of charge, the magnitude of the charge on an electron. It follows that the charge on an ion is denoted by the integer number z of the fundamental unit of charge, and the mass-to-charge ratio m/z therefore represents Daltons per fundamental unit of charge. In many cases, the ions encountered in mass spectrometry have just one charge (z=1) so the m/z value is numerically equal to the molecular (ionic) mass in Da.

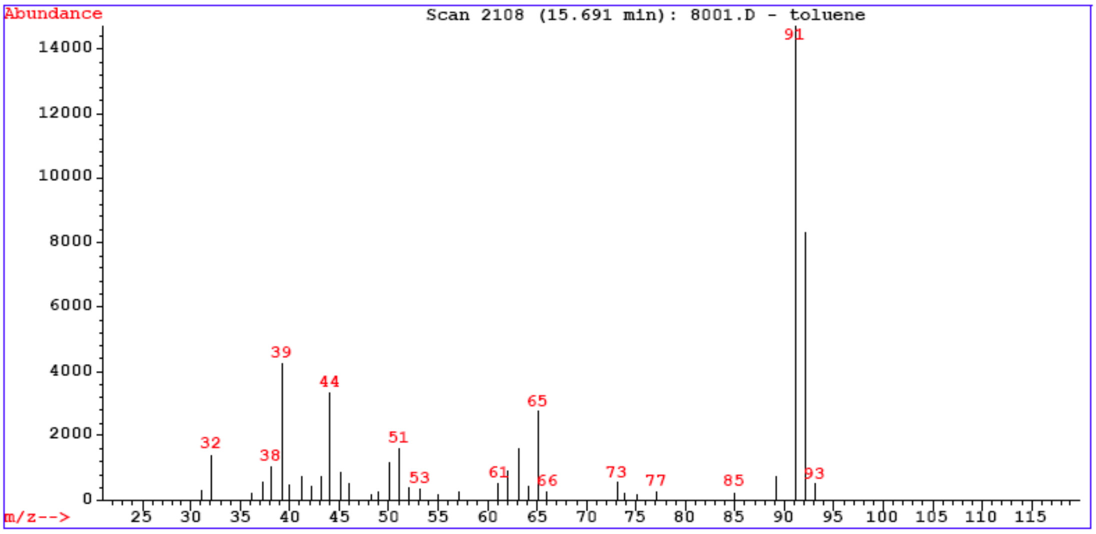


FIGURE 2-5 THE TOLUENE MASS SPECTRUM. ON THE Y COORDINATE WE HAVE THE COUNTS OF IONS (I.E. INTENSITY) AND ON THE X COORDINATE THERE ARE THE M/Z RATIO [Da].

All mass spectrometers consist of three basic parts: an ion source, a mass analyzer, and a detector system (see Figure 2-6). The stages within the mass spectrometer are:

1. Producing ions from the sample (ionization source);
2. Separating ions based on mass-to-charge ratio (mass analyzer);
3. Detecting the number of ions of each mass produced (detector);
4. Collecting, processing and analyzing the results and generating the mass spectrum (data system).

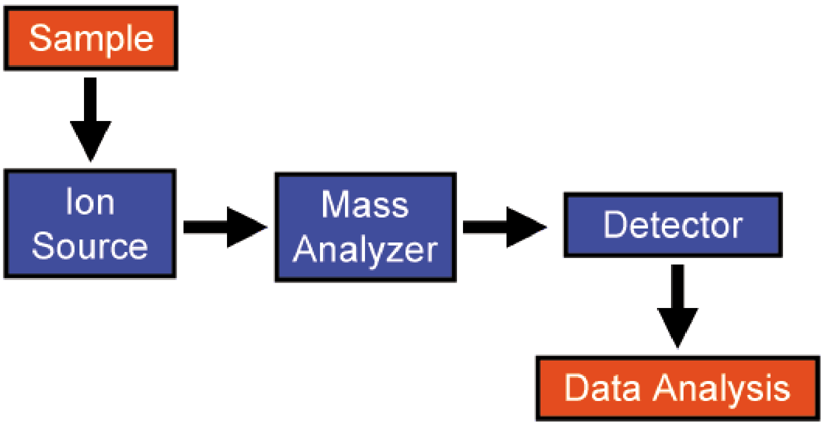


FIGURE 2-6 SCHEME OF THE FUNCTIONAL BLOCKS OF A MASS SPECTROMETER.

ıtion

source. The sample, which may be a solid, liquid, or vapor, enters the vacuum chamber of the MS through an inlet. Depending on the type of inlet and ionization techniques used, the sample may already exist as ions in solution, or it may be ionized in conjunction with its volatilization or by other methods in the ion source. After ionizing the sample, the ions of the sample are passed to the mass analyzer region where separation based on the mass-to-charge ratio occurs. Once separated by the analyzer, the ions then enter the detector portion of the mass spectrometer. At this point, the machine calculates the mass-to-charge ratio and the relative abundance of each of the different ions. From this information, a spectrum graph can be created. Most mass spectrometers are maintained under a vacuum to improve the chances of ions traveling from ionization source to detector without interference by collision with air molecules.

2.3.1 THE IONIZATION SOURCE

The ion source is the mass spectrometer component which ionizes the sample to be analyzed. Ionization mainly serves to present the sample as vaporized ions which can be acted upon by the mass analyzer and measured by the ion detector. Formation of gas phase samples ions is an essential prerequisite to the mass sorting and detection processes that occur in a mass spectrometer.

There are many different methods available to ionize samples, such as positive or negative ion modes. The ionization method chosen should depend on the type of sample and the type of mass spectrometer. There are two main classes of ionization methods, electron and chemical. Electron ionization involves application of an electrical current to the sample to induce ionization. Chemical ionization involves interaction of the sample with reagent molecules to induce ionization. Ions produced are often denoted with symbols that indicate the nature of the ionization: for example, $[M+H]^+$ is used to represent a molecule which is protonated.

The development of new ionization sources has been pivotal for the application of MS to biological samples and, therefore, the birth of the MS-based proteomics. Early mass spectrometers required a sample to be a gas: this was a great limit for its applicability to biological samples. In 2002, the Nobel Prize in Chemistry was received by John Bennett Fenn for the development of a soft ionization technique for liquid solutions, electrospray ionization (ESI) (see Figure 2-7), and Koichi Tanaka for the development of soft laser desorption (SLD) in 1987. An improved SLD method, matrix-assisted laser desorption/ionization (MALDI), was developed in 1987 by Franz Hillenkamp and Michael Karas. ESI and MALDI made it possible to apply mass spectrometry to samples in liquid solutions or embedded in a solid matrix.

In particular, soft ionization techniques were pivotal for proteomics research. "Soft" in the context of ion formation means forming ions without breaking chemical bonds. Indeed, in biological studies where the analyst often requires that non-covalent molecule-protein or protein-protein interactions are representatively transferred into the gas-phase, the formation of gas-phase ions without extensive fragmentation is mandatory. Two soft ionization methods commonly used in proteomics are ‘Matrix Assisted Laser Desorption Ionization’ or MALDI and ‘Electrospray Ionization’ also known as ESI.

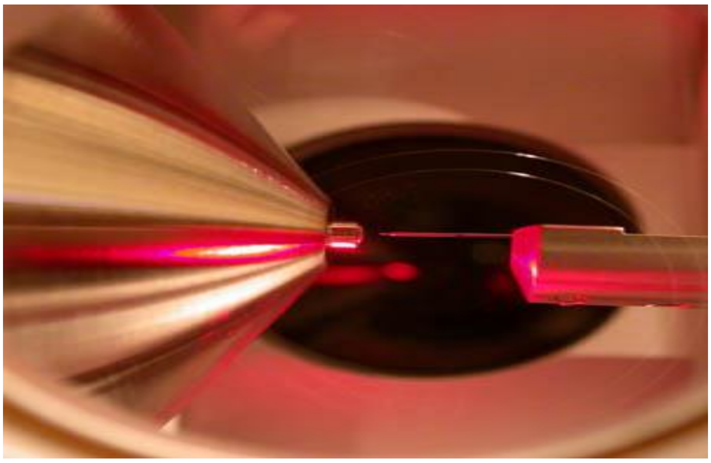


FIGURE 2-7 A NANO-ELECTROSPRAY ION SOURCE (NANO-ESI).

2.3.1.1 ELECTROSPRAY IONIZATION

Electrospray ionization (ESI) is a very popular electron ionization technique in mass spectroscopy for ionizing samples before they are measured. ESI works well with heavier compounds and is therefore often used in proteomics. In particular, it overcomes the propensity of these molecules to fragment when ionized. Electrospray can be simply considered an interface for transferring ions from the solution phase to the gas phase. The development of electrospray ionization for the analysis of biological macromolecules was rewarded with the attribution of the Nobel Prize in Chemistry to John Bennett Fenn in 2002 [6,7].

The ESI source has undergone continued development since the earliest examples, but the general arrangement, as reported in Figure 2-8, has remained basically the same.

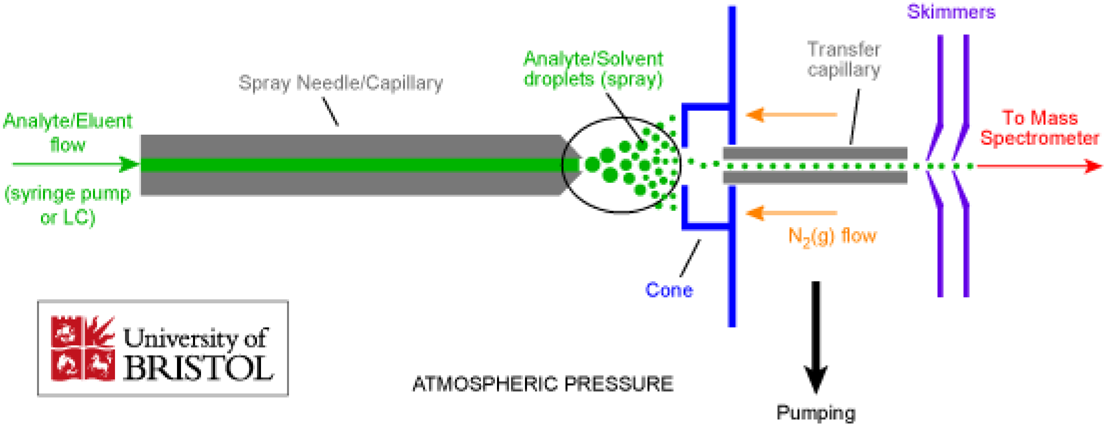


FIGURE 2-8 A SCHEME REPRESENTING THE ESI WORKFLOW. ESI IS AN ATMOSPHERIC PRESSURE IONIZATION TECHNIQUE. IONS ARE FORMED IN SOLUTION (DROPLETS) AND THEN THE DROPLETS ARE EVAPORATED WITH A DRYING GAS (NEBULISED) IN THE PRESENCE OF A STRONG ELECTROSTATIC FIELD. THIS WILL DISASSOCIATE MOLECULES , INCREASE THE CHARGE CONCENTRATION. EVENTUALLY THE REPULSIVE FORCE BETWEEN IONS WITH LIKE CHARGES EXCEEDS THE COHESIVE FORCES AND IONS ARE EJECTED IN TO THE GAS PHASE.

The analyte is introduced to the source in solution either from a syringe pump or as the eluent flow from liquid chromatography. The analyte solution flow passes through an electrospray needle, where a high potential difference is applied with respect to the counter electrode (from 2.5 to 4 kV). This forces the spraying of charged droplets from the needle with a surface charge of the same polarity to the charge on the needle. Since the droplets have the same charge, they are repelled from the needle towards the source sampling cone on the counter electrode. As the droplets traverse the space between the needle tip and the cone and solvent evaporation occurs. As the solvent evaporation occurs, the droplet shrinks until it reaches the point that the surface tension can no longer sustain the charge (the Rayleigh limit) at which point a "Coulombic explosion" occurs and the droplet is ripped apart. This produces smaller droplets that can repeat the process as well as naked charged analyte molecules. These charged analyte molecules (they are not strictly ions) can be singly or multiply charged.

ESI is a very soft method of ionization as very little residual energy is retained by the analyte upon ionization (see Figure 2-8). This is why ESI-MS is such an important technique in proteomics.

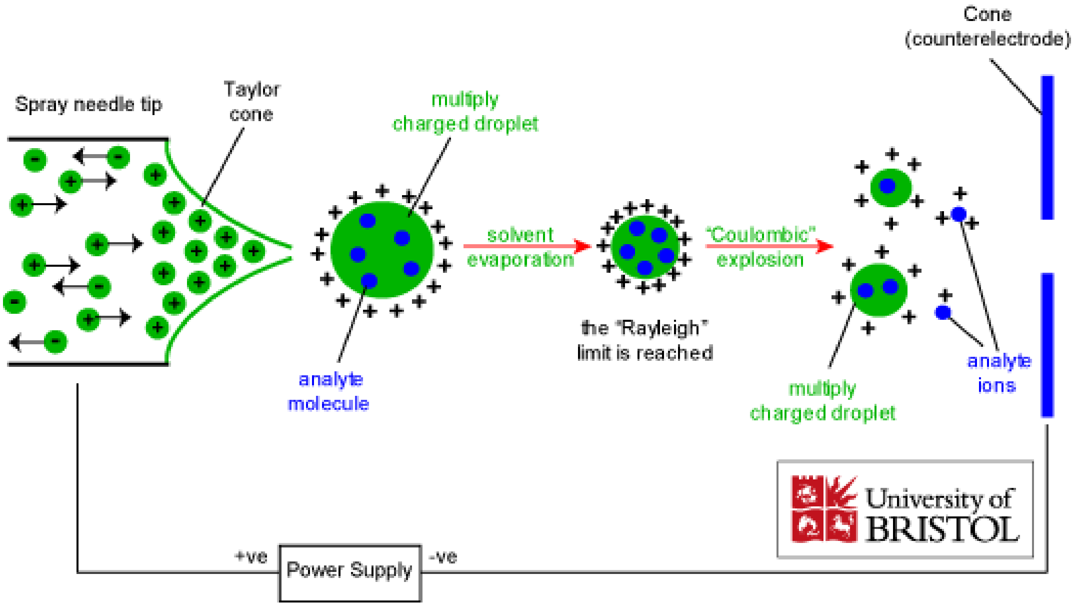


FIGURE 2-9 A SCHEME REPRESENTING THE ION FORMATION IN ESI.

There are many variations on the basic electrospray technique, that generally offer better sensitivity than it. Two important ones are microspray (µ-spray) and nanospray. The primary difference is in the reduced flow rate of the analyte containing liquid, µLiters/minute and nLiters/minute respectively; this causes many other differences, such as the reduced internal diameter of the tubing or lack of nebulization gas.

2.3.1.2 MATRIX-ASSISTED LASER DESORPTION IONIZATION

Matrix-assisted laser desorption ionization (MALDI) is a soft ionization technique used in mass spectrometry, allowing the analysis of biomolecules (biopolymers such as proteins, peptides and sugars) and large organic molecules (such as polymers and other macromolecules), which tend to be fragile and fragment when ionized by more conventional ionization methods. It is most similar in character to electrospray ionization both in relative softness and the ions produced, although MALDI causes much fewer multiply charged ions. Most ions are found in the +1 charge state $[M+H]^+$. The ionization is triggered by a laser beam (normally a nitrogen laser). A matrix is used to protect the biomolecules from being destroyed by direct laser beam and to facilitate vaporization and ionization.

The term matrix-assisted laser desorption ionization (MALDI) was coined in 1985 by Franz Hillenkamp, Michael Karas and their colleagues [8,9]. The breakthrough for large molecule laser desorption ionization came in 1987 when Koichi Tanaka of Shimadzu Corp. and his co-workers used what they called the “ultra-fine metal plus liquid matrix method” [10]. Tanaka received one-quarter of the 2002 Nobel Prize in Chemistry for demonstrating that, with the proper combination of laser wavelength and matrix, a protein can be ionized. The availability of small and relatively inexpensive nitrogen lasers operating at 337 nm wavelength and the first commercial instruments introduced in the early 1990s brought MALDI to an increasing number of researchers.

The identity of suitable matrix compounds is determined to some extent by trial and error, but they are based on some specific molecular design considerations. They are of a fairly low molecular weight, to facilitate vaporization, but are large enough, with a high enough vapor pressure, not to evaporate during sample preparation or while standing in the spectrometer. They are acidic, therefore act as a proton source to encourage ionization of the analyte. They have a strong optical absorption in the UV, so that they rapidly and efficiently absorb the laser irradiation. They are functionalized with polar groups, allowing their use in aqueous solutions. The matrix solution is mixed with the analyte (e.g. protein-sample): the organic solvent allows hydrophobic molecules to dissolve into the solution, while the water allows for water-soluble (hydrophilic) molecules to do the same. This solution is spotted onto a MALDI plate that usually is a metal plate designed for this purpose (see Figure 2-10). The solvents vaporize, leaving only the re-crystallized matrix, but now with analyte molecules spread throughout the crystals. Thus the matrix and the analyte are said to be co-crystallized in a MALDI spot.

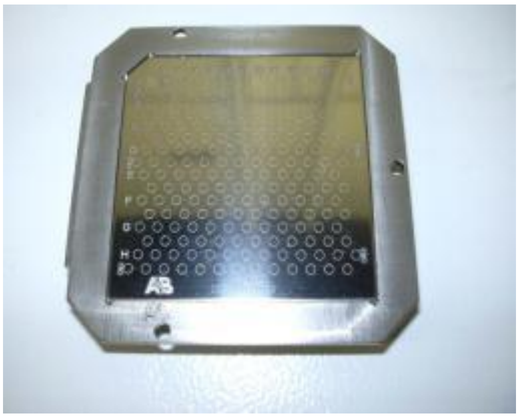


FIGURE 2-10 SAMPLE TARGET FOR MALDI.

The laser hits the spot on the crystallized matrix and transfers energy from the matrix molecule to the sample. This energy transfer vaporizes the sample, sending a plume of ions into the MALDI source. This plume of ions is then collected and held in the source until a pulse sends them all out simultaneously (see Figure 2-11). If the MALDI is attached to a Time of Flight (TOF) mass analyzer these ions are then sent down the TOF tube and are separated according to their velocity (light ions hitting first). The TOF mass analyzer will be described in the following of this chapter.

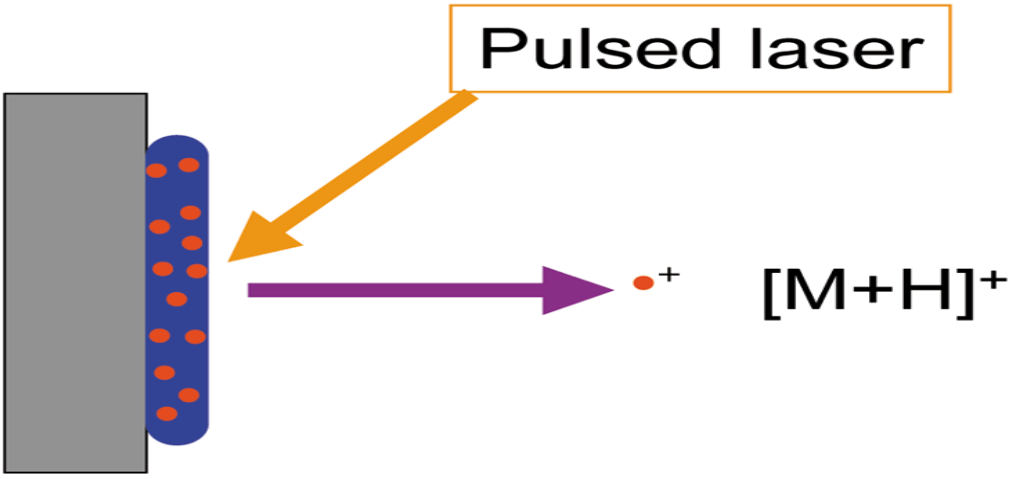


FIGURE 2-11 SCHEMATIC REPRESENTATION OF HOW MALDI IONIZE THE SAMPLE. MOST OF THE IONS ARE FOUND IN THE +1 CHARGE STATE $[M+H]^+$.

2.3.1.3 SURFACE-ENHANCED LASER DESORPTION IONIZATION

Surface-enhanced laser desorption ionization (SELDI) is a variation of MALDI that uses a target modified to achieve biochemical affinity with the analyte compound [11]. In MALDI, a protein or peptide sample is mixed with the matrix molecule in solution and small amounts of the mixture are deposited on a surface and allowed to dry. The sample and matrix co-crystallize as the solvent evaporates. In SELDI the protein mixture is spotted on a surface modified with a chemical functionality. Some proteins in the sample bind to the surface, while the others are removed by washing. After washing the spotted sample, the matrix is applied to the surface and allowed to crystallize with the sample peptides. Binding to the SELDI surface acts as a separation step. The subset of proteins that binds to the surface are easier to analyze. Common surfaces include weak-positive ion exchange, hydrophobic surface, metal-binding surface, strong anion exchanger. Surfaces can also be functionalized with antibodies, other proteins, or DNA.

SELDI is used to detect proteins in tissue samples, blood, urine, or other clinical samples.

Samples spotted on a SELDI surface are typically analyzed using the TOF mass analyzer. A laser ionizes peptides from crystals of the sample/matrix mixture. The ions are accelerated through an electric potential and down a flight tube. A detector measures ions as they reach the end of the tube. The mass-to-charge ratio of each ion can be determined from the length of the tube, the kinetic energy given to ions by the electric field, and the time taken to travel the length of the tube.

2.3.2 THE MASS ANALYZER

The analyzer uses dispersion or filtering to sort ions according to their mass-to-charge ratios or a related property. The most widely used analyzers are sectors, quadrupole mass filters, quadrupole ion traps, Fourier transform ion cyclotron resonance spectrometers, and time-of-flight mass analyzers.

Mass analyzers separate the ions according to their mass-to-charge ratio. All mass spectrometers are based on dynamics of charged particles in electric and magnetic fields in vacuum where the Lorentz’s force law (2-1) and the Newton's second law of motion (2-2) apply:

$$\mathbf{F} = q\,(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \qquad 2\text{-}1$$

$$\mathbf{F} = m\mathbf{a} \qquad 2\text{-}2$$

where **F** is the force applied to the ion, m is the mass of the ion, $\mathbf{a}=\dot{\mathbf{v}}$ is the acceleration, q is the ionic charge, **E** is the electric field, and **v** x **B** is the vector cross product of the ion velocity and the magnetic field. Equating the above expressions for the force applied to the ion yields:

$$(m/q)\,\mathbf{a} = \mathbf{E} + \mathbf{v} \times \mathbf{B} \qquad 2\text{-}3$$

This differential equation 2-3 is the classic equation of motion of charged particles. Together with the particle's initial conditions it completely determines the particle's motion in space and time and therefore is the basis of every mass spectrometer. It immediately reveals that two particles with the same physical quantity m/q behave exactly the same. So what equation (2-3) is basically saying is that the mass to charge ratio acts as a determinant of acceleration of the ion, which can also be represented as the addition of the electric field plus the cross product of the ion velocity and magnetic field.

2.3.2.1 SECTOR FIELD MASS ANALYZER

A sector field mass analyzer (see Figure 2-12) uses an electric and/or magnetic field to affect the path and/or velocity of the charged particles: it changes the direction of ions that are accelerated through the mass analyzer. The ions enter a magnetic or electric field which bends the ion paths depending on their mass-to-charge ratios, deflecting the more charged and faster, lighter ions. Under ideal conditions ions of different masses will separate physically in space into different beams. Ions of larger m/z follow larger radius paths than ions of smaller m/z values so ions of differing m/z values are dispersed in space. By changing the ion trajectories through variations of the magnetic field strength, ions of different nominal mass-to-charge ratios can be focused on a detector. The ions eventually reach the detector and their relative abundances are measured. The analyzer can be used to select a narrow range of m/q or to scan through a range of m/q to catalog the ions present. Double focusing mass spectrometers use a combination of magnetic and electrical fields to focus and sort ions.

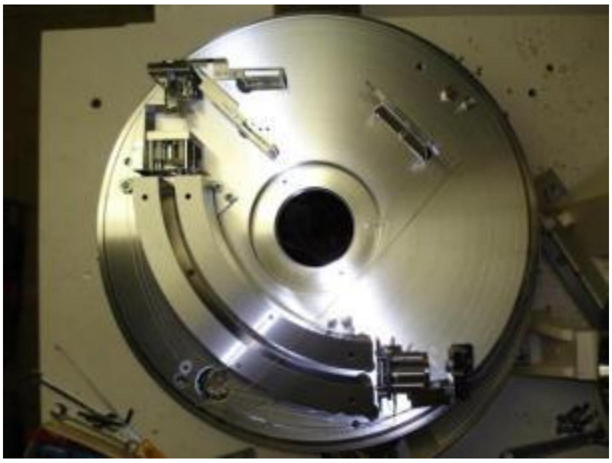


FIGURE 2-12 A SECTOR FIELD FROM A FINNIGAN MAT MASS SPECTROMETER.

2.3.2.2 TIME-OF-FLIGHT MASS ANALYZER

The Time-of-flight (TOF) is a mass analyzer that allows ions to flow down a field free region [12]. This allows the ions with a greater velocity, lighter ions, to hit the detector first. TOF is especially compatible with MALDI (or SELDI) due to the fact that it needs a pulsed source for ions emission. Ions are generated in the MALDI source and then all are pulsed into the TOF at the same exact time. This results in all the ions receiving the same initial kinetic energy. Therefore, the ions with the lower mass will have a higher velocity and reach the detector first; whereas the ions with the higher mass will have slower velocity and hit the detector last. The time that it takes for the particle to reach the detector at a known distance is measured and it is the so called time of flight. It will depend on the mass-to-charge ratio of the particle (heavier particles reach lower speeds). From the time of flight and the known experimental parameters it is possible to compute the mass-to-charge ratio of the ion.

2.3.2.3 FOURIER TRANSFORM ION CYCLOTRON RESONANCE MASS ANALYZER

Fourier transform ion cyclotron resonance (FTICR) is a type of mass analyzer which determines the mass-to-charge ratio of ions based on the cyclotron frequency of the ions in a fixed magnetic field [13]. The ions are trapped in a Penning trap (a magnetic field with electric trapping plates) where they are excited to a larger cyclotron radius by an oscillating electric field perpendicular to the magnetic field. The excitation also results in the ions moving in phase, so you can imagine them moving like in packets. The signal is detected as an image current on a pair of plates which the packet of ions passes close to as they cycle around. The resulting signal is called a free induction decay (FID) transient or interferogram and consists of a superposition of sine waves. The useful signal is extracted from this data using the properties of the Fourier transform in order to obtain a mass spectrum. FTMS has the advantage of high sensitivity (since each ion is 'counted' more than once) and much high resolution and thus precision.

2.3.2.4 THE ORBITRAP

One of the most recently introduced mass analyzers is the Orbitrap [14-17]. Here, ions are electrostatically trapped in an orbit around a central, spindle-shaped electrode. Ions are injected tangentially into the electric field between the electrodes and trapped because their electrostatic attraction to the inner electrode is balanced by centrifugal forces. Thus, ions cycle around the central electrode in rings. In addition, the ions also move back and forth along the axis of the central electrode. Therefore, ions of a specific mass-to-charge ratio move in rings which oscillate along the central spindle. The frequency of these harmonic oscillations is independent of the ion velocity and is inversely proportional to the square root of the mass-to-charge ratio. This oscillation generates a current in the detector plates which is recorded by the instrument. The frequencies of these currents depend on the mass to charge ratios of the ions in the Orbitrap. Mass spectra are obtained by Fourier transformation of the recorded image currents. Orbitraps have a high mass accuracy (1-2 ppm), a high resolving power (up to 200,000) and a high dynamic range (around 5000) [18]. Like FTICR-MS the Orbitrap resolving power is proportional to the number of harmonic oscillations of the ions, as a result the resolving power is inversely proportional to the square root of m/z and proportional to acquisition time. Given that a transient is the duration that the time domain signal is acquired for, the resolving power decreases further as the m/z value increases so that at 4 times the m/z value the resolving power has halved. Approximately 0.1 seconds per transient is required for data processing, thus a 0.1 second transient has a cycle time of 0.2 seconds.

Further improvements on the Orbitrap technology have been achieved during last decade. OrbitrapXL and the newest Orbitrap Velos feature faster acquisitions, higher resolutions and accuracies than their common ancestor.

2.3.2.5 THE QUADRUPOLE MASS ANALYZER

The quadrupole mass analyzer [19] is essentially a mass filter that is capable of transmitting only the ion of choice. A mass spectrum is obtained by scanning through the mass range of interest over time. The quadrupole consists of four parallel metal rods. Each opposing rod pair is connected together electrically and a radio frequency (RF) voltage is applied between one pair of rods and the other. A direct current voltage is then superimposed on the RF voltage. Ions travel down the quadrupole in between the rods. Only ions of a certain m/z will reach the detector for a given ratio of voltages: other ions have unstable trajectories and will collide with the rods. This allows selection of a particular ion, or scanning by varying the voltages and thus the selected ions.

These types of mass spectrometers excel at applications where particular ions of interest are studied because they can stay tuned on a single ion for extended periods of time. One place where this is useful is in liquid chromatography-mass spectrometry or gas chromatography-mass spectrometry where they serve as exceptionally high specificity detectors. Quadrupole instruments are often reasonably priced and make good multi-purpose instruments, but they provide lower resolution than double focusing instruments.

2.3.2.6 THE QUADRUPOLE ION TRAP MASS ANALYZER

A quadrupole ion trap [20] exists in both linear and 3D (Paul Trap, QIT) varieties and refers to an ion trap that uses DC (direct current) and radio frequency (RF) oscillating AC (alternating current) electric fields to trap ions. The invention of the 3D quadrupole ion trap itself is attributed to Wolfgang Paul [21] who shared the Nobel Prize in Physics in 1989 for this work.

The quadrupole ion trap mass spectrometer (see Figure 2-13) operates on a principle similar to a quadrupole mass filter. However, it does not operate as a filter. Rather, the ion trap stores ions for subsequent experiments and analysis. It uses fields generated by RF (and sometimes DC) voltages applied to electrodes arranged in a sandwich geometry. The 3D trap itself generally consists of two hyperbolic metal electrodes with their focuses facing each other and a hyperbolic ring electrode halfway between the other two electrodes. The ions are trapped in the space between these three electrodes by AC (oscillating, non-static) and DC (non-oscillating, static) electric fields. The AC radio frequency voltage oscillates between the two hyperbolic metal end cap electrodes if ion excitation is desired; the driving AC voltage is applied to the ring electrode. The ions are first pulled up and down axially while being pushed in radially.

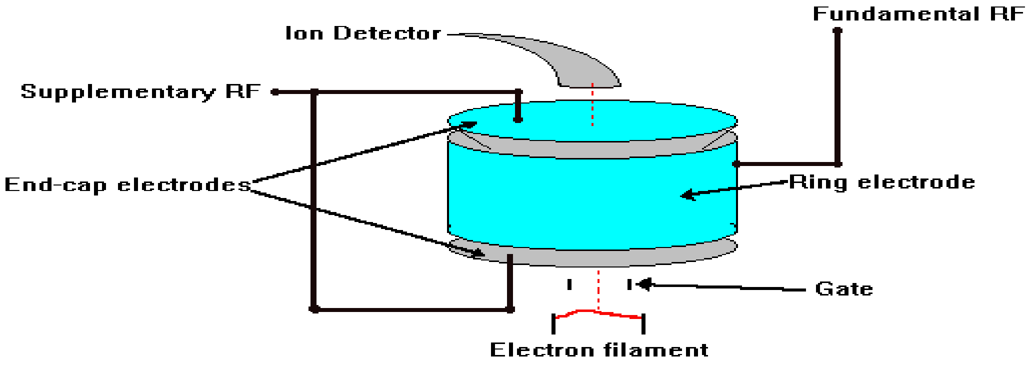
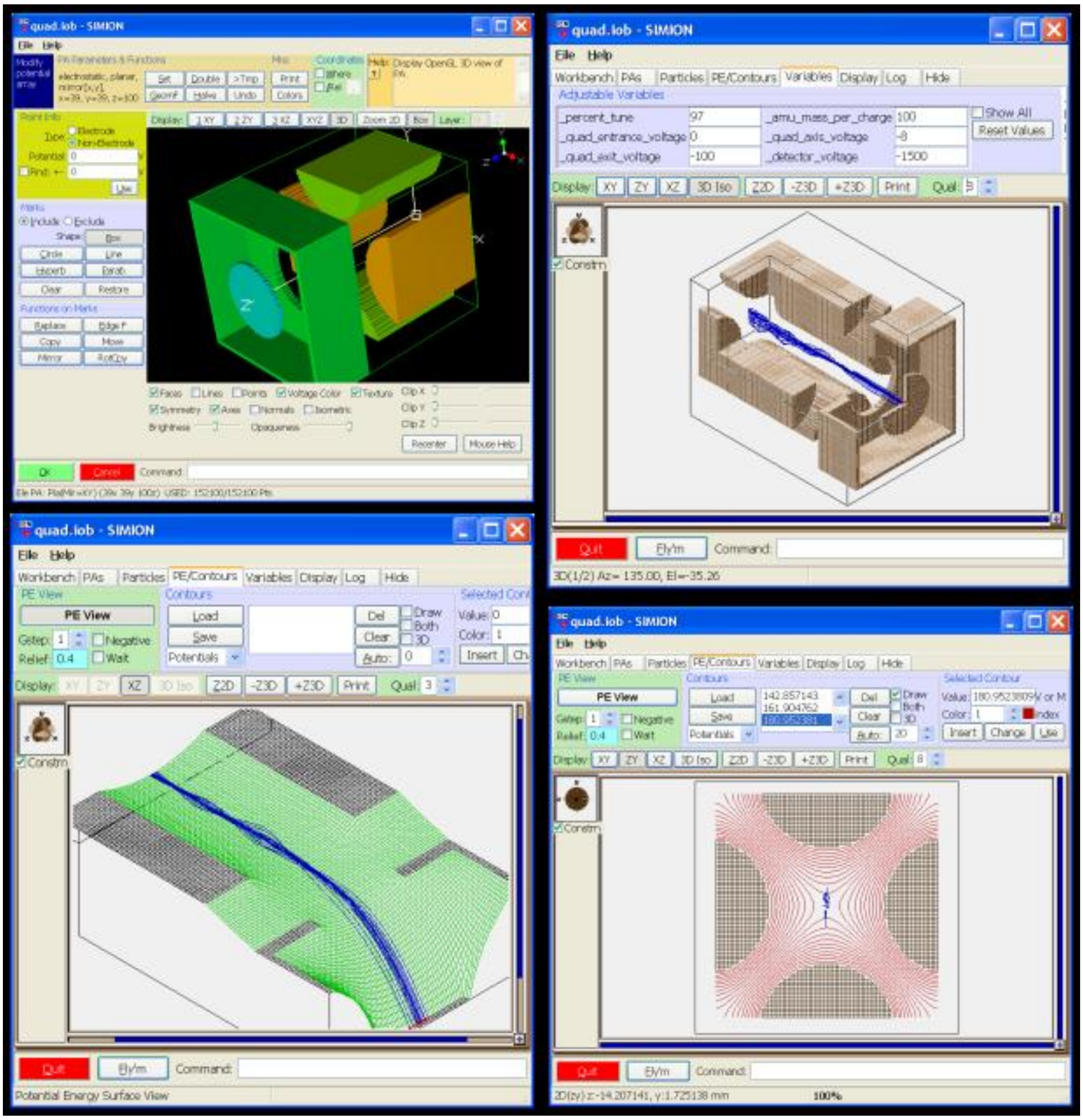


FIGURE 2-13 A SCHEME REPRESENTATION OF A QIT.

The ions are then pulled out radially and pushed in axially (from the top and bottom). In this way the ions move in a complex motion that generally involves the cloud of ions being long and narrow and then short and wide, back and forth, oscillating between the two states (see Figure 2-14). The quadrupole ion trap has two configurations: the three dimensional form described above and the linear form made of 4 parallel electrodes.



_percent_tune	97	_amu_mass_per_charge	100
_quad_entrance_voltage	0	_quad_axis_voltage	-8
_quad_exit_voltage	-100	_detector_voltage	-1500

FIGURE 2-14 IN THE LEFT UPPER IMAGE THERE IS A 3D VISUALIZATION OF A QIT. IN THE RIGHT UPPER IMAGE YOU CAN SEE THE IONS TRAJECTORY THROUGHOUT THE QUADRUPOLE. BELOW, ON THE LEFT YOU SEE THE REPRESENTATION OF THE POTENTIAL ENERGY SURFACE, WHILE ON THE RIGHT THERE IS A DEPICTION OF THE ELECTROMAGNETIC FIELD IN THE TRAP (THOSE PINK LINES YOU SEE ARE THE EQUIPOTENTIAL LINES IN THE TRAP).

A linear quadrupole ion trap (LTQ) (see Figure 2-15) is similar to a QIT, but traps ions in a 2D quadrupole field, instead of a 3D quadrupole field as in a QIT. Linear ion trap uses a set of quadrupole rods to confine ions radially and a static electrical potential on end electrodes to confine the ions axially. The linear form of the trap can be used as a selective mass filter or as an actual trap by creating a potential well for the ions along the axis of the electrodes. Advantages of the linear trap design are increased ion storage capacity, faster scan times, and simplicity of construction, although quadrupole rod alignment is critical, adding a quality control constraint to their production. LTQ is the quadrupole used to generate our dataset.

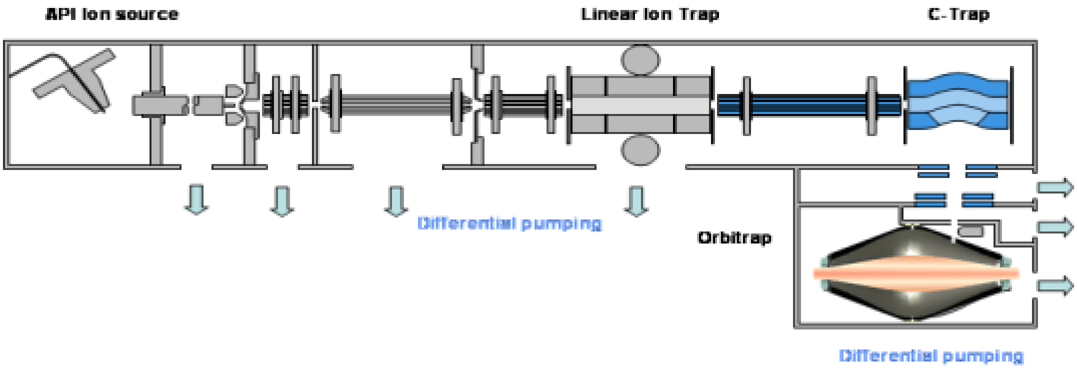


FIGURE 2-15 A LTQ SCHEME.

The subsequent Figure 2-16, borrowed from [3], graphically summarizes the main kind of mass spectrometers used nowadays in proteome research.

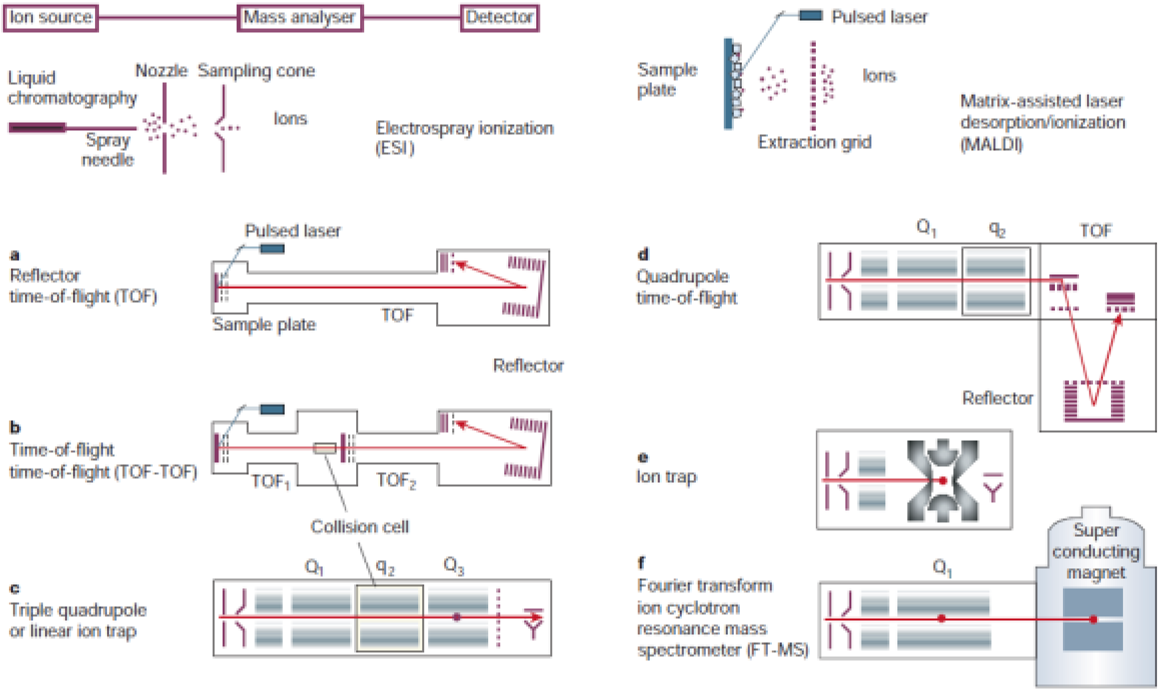

Ion source
Mass analyser
Detector
Liquid chromatography
Nozzle
Sampling cone
Ions
Spray needle
Electrospray ionization (ESI)
Sample plate
Pulsed laser
Ions
Extraction grid
Matrix-assisted laser desorption/ionization (MALDI)
a
Reflector time-of-flight (TOF)
Pulsed laser
Sample plate
TOF
Reflector
b
Time-of-flight time-of-flight (TOF-TOF)
TOF1
TOF2
Collision cell
c
Triple quadrupole or linear ion trap
Q1
q2
Q3
d
Quadrupole time-of-flight
Q1
q2
TOF
Reflector
e
Ion trap
f
Fourier transform ion cyclotron resonance mass spectrometer (FT-MS)
Q1
Super conducting magnet

FIGURE 2-16 MASS SPECTROMETERS USED NOWADAYS IN PROTEOME RESEARCH.

2.3.3 THE DETECTOR

The final element of the mass spectrometer is the detector: the ions which pass through the analyzer are now separated by the desired methods. The detector records the charge induced or current produced when an ion passes by or hits a surface. In a scanning instrument the signal produced in the detector during the course of the scan versus where the instrument is in the scan (at what m/q) will produce a mass spectrum, a record of ions as a function of m/q. Due to the fact that the number of ions entering the detector at any given moment is extremely small, signal amplification is often necessary. Typically, some type of electron multiplier is used, though other detectors including Faraday cups and ion-to-photon detectors are also used. Microchannel Plate Detectors are commonly used in modern commercial instruments. In FTICR-MS and Orbitrap, the detector consists of a pair of metal surfaces within the mass analyzer/ion trap region which the ions only pass near as they oscillate. No DC current is produced, only a weak AC image current is produced in a circuit between the electrodes. In the following the most used detectors will be described.

2.3.3.1 THE FARADAY CUP

A faraday cup is a metal (i.e. conductive) cup designed to catch charged particles in vacuum. The resulting current can be measured and used to determine the number of ions or electrons hitting the cup. The Faraday cup (see Figure 2-17) is named after Michael Faraday who first theorized ions around 1830. When a beam or packet of ions hits the metal it gains a small net charge while the ions are neutralized. The metal can then be discharged to measure a small current equivalent to the number of impinging ions. Essentially the faraday cup is part of a circuit where ions are the charge carriers in vacuum and the faraday cup is the interface to the solid metal where electrons act as the charge carriers (as in most circuits). By measuring the electrical current (the number of electrons flowing through the circuit per second) in the metal part of the circuit the number of charges being carried by the ions in the vacuum part of the circuit can be determined. Faraday cups are not as sensitive as electron multiplier detectors, but are highly regarded for accuracy because of the direct relation between the measured current and number of ions.

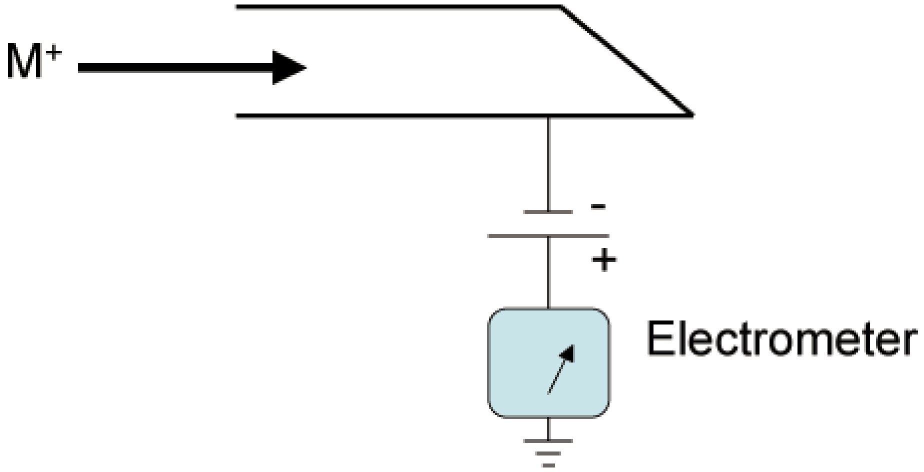


FIGURE 2-17 A FARADAY CUP SCHEME.

2.3.3.2 ELECTRON MULTIPLIERS

An electron multiplier (continuous dynode electron multiplier) is a vacuum-tube structure that multiplies incident charges. In a process called secondary emission, a single electron can, when bombarded on metal (or PbO coated surface) induce emission of roughly 1 to 3 electrons. If an electric potential is applied between this metal plate and yet another, the emitted electrons will accelerate to the next metal plate and induce secondary emission of still more electrons. This can be repeated a number of times, resulting in a large shower of electrons all collected by a metal anode, all having been triggered by just one. Therefore, another name for electron multipliers is avalanching ion detector: 12 stages of acceleration will usually give a gain in current of 10 million electrons. The avalanche can be triggered by any charged particle hitting the starting electrode with sufficient energy to cause secondary emission. It could also be triggered by a photon causing vacuum photoemission of at least one electron. In a photomultiplier tube (see Figure 2-18), a photo-emissive surface is followed by an electron multiplier with several sequential multiplying electrodes called dynodes.

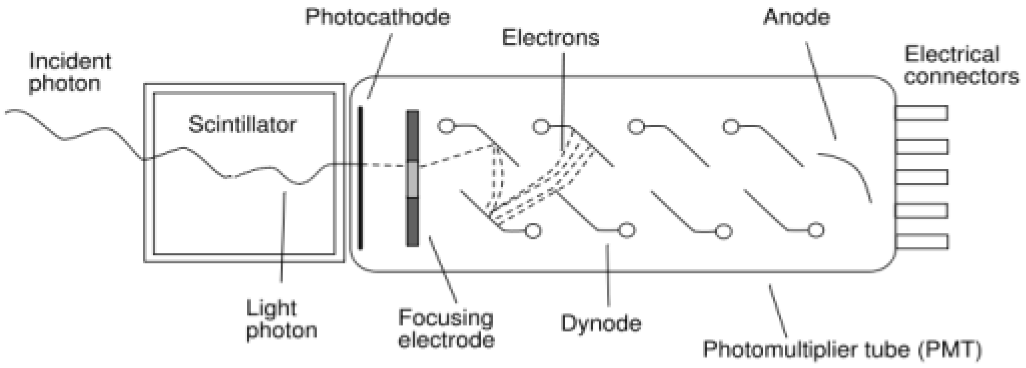


FIGURE 2-18 A PHOTOMULTIPLIER TUBE SCHEME.

Because these electrodes are separate from each other, this might be called a "discrete-dynode" multiplier. A voltage divider chain of resistors is usually used to place each dynode at a potential 100-200V more positive than the previous one. A "continuous-dynode" structure is feasible if the material of the electrodes has a high resistance, so that the functions of secondary-emission and voltage-division are merged; this is often built as a funnel of glass coated inside with a thin film of semi-conducting material, with negative high voltage applied at the wider input end, and positive voltage near ground applied at the narrower output end. Electrons emitted at any point are accelerated a modest distance down the funnel before impacting the surface, perhaps on the opposite side of the funnel. At the destination end a separate electrode (anode) remains necessary to collect the multiplied electrons. In mass spectrometry electron multipliers are often used as a detector of ions that have been separated by a mass analyzer of some sort. They are typically of the continuous-dynode type, and may have a curved horn-like funnel shape.

2.3.3.3 MICRO-CHANNEL PLATE DETECTORS

A micro-channel plate (MCP) (see Figure 2-19) is a planar component used for detection of particles (electrons or ions) and impinging radiation (ultraviolet radiation and X-rays). It is closely related to an electron multiplier, as both intensify single particles or photons by the multiplication of electrons via secondary emission. Each microchannel is a continuous-dynode electron multiplier, in which the multiplication takes place under the presence of a strong electric field. A particle or photon that enters one of the channels through a small orifice is guaranteed to hit the wall of the channel due to the channel being at an angle to the plate and thus the angle of impact. The impact starts a cascade of electrons that propagates through the channel, which amplifies the original signal by several orders of magnitude depending on the electric field strength and the geometry of the micro-channel plate. After the cascade, the microchannel takes time to recover (or recharge) before it can detect another signal. The electrons exit the channels on the opposite side where they are themselves detected by additional means, often simply a single metal anode measuring total current. In some applications each channel is monitored independently to produce an image. Phosphors in combination with photomultiplier tubes have also been used.

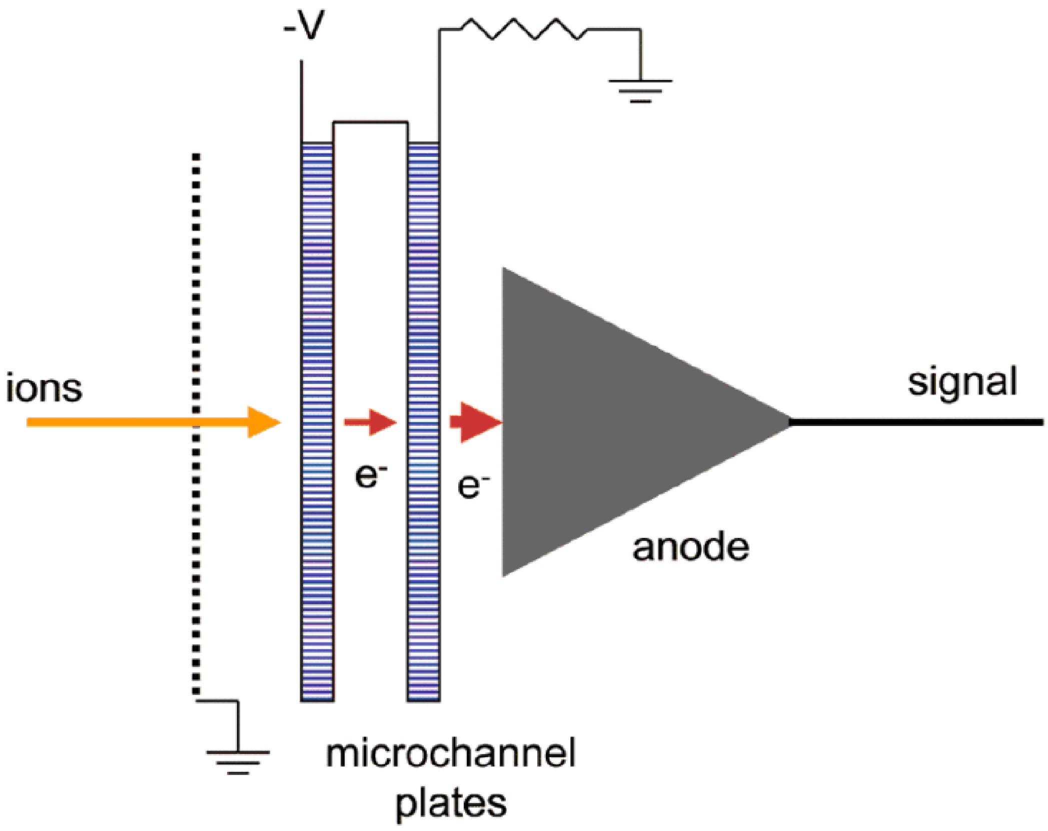


FIGURE 2-19 DUAL MICROCHANNEL PLATE DETECTOR SCHEMATIC.

2.4 COMMON MASS SPECTROMETER CONFIGURATIONS

When all of the elements (source, analyzer and detector) of a mass spectrometer are combined to form a complete instrument and the specific configuration becomes common a new name, often an abbreviation of one or more of the internal components, becomes attached to the specific configuration and can become more well-known than the specific internal components. Sometimes the use of the generic "MS" actually implies a very specific mass analyzer and detection system, which is always sector based. In other cases there are common configurations that may be implied but not necessarily. An important enhancement to the mass resolving and determining capacity of mass spectrometry is the combination of mass spectrometry with analysis techniques that resolve mixtures of compounds in a sample based on other characteristics before introduction into the mass spectrometer.

2.4.1 MALDI-MS

Matrix Assisted Laser Desorption Ionization mass spectrometry (MALDI-MS) [9] deals with thermo labile, non-volatile organic compounds and those of high molecular mass. It is used in for the analysis of proteins, peptides, glycoproteins, oligosaccharides, and oligonucleotides.

MALDI is based on the usage of matrix complexed with a given sample molecule that is bombarded with a laser in order for the sample molecule to form a sample ionization. The sample is normally mixed into a high absorbable matrix with as little matrix as possible as the matrix will also become excited and come off and ionize as well. The matrix itself acts as a substance which infuses the sample as well as a transformer for the laser's energy into excitation energy to allow for the vaporization of the sample ions and matrix ions from the surface of the matrix. Most commercially available MALDI mass spectrometers are now a pulsed nitrogen laser of wavelength 337 nm. In order to obtain proper charge-mass ratios and calculate a mass spectrum the type of mass spectrometer most widely used with MALDI is the TOF (time-of-flight mass spectrometer), mainly due to its large mass range. The TOF measurement procedure is also ideally suited to the MALDI ionization process since the pulsed laser takes individual 'shots' rather than working in continuous operation. MALDI-TOF instruments (see Figure 2-20) are typically equipped with an "ion mirror", deflecting ions with an electric field, thereby doubling the ion flight path and increasing the resolution. Commercial reflectron TOF instruments reach today a resolving power m/Δm of well above 20'000 FWHM (full-width half-maximum, Δm defined as the peak width at 50% of peak height).

FIGURE 2-20 A MALDI-TOF INSTRUMENT.

2.4.2 SELDI-MS

Surface Enhanced Laser Desorption Ionization mass spectrometry is a modification of the procedure used in MALDI-MS. Instead of mixing the UV sensitive matrix with the protein sample, the protein sample is spotted on a plate which has some surface binding characteristics such as a chromatographic array. The target surfaces, to which the proteins and matrices are applied to, are coated with various activated and patented chemistries. Therefore, it is possible to fractionate proteins within a mixture, or particular classes of proteins, on the chip or array surface prior to analysis. The spots are then washed to remove impurities and weakly bound proteins. The UV matrix is then added to the spot and allowed to co-crystallize. After the ionization with the UV laser, the ions are analyzed using a TOF mass analyzer, in the same manner as MALDI. The reason for fractionating samples prior to analysis is not only to make the analysis much simpler but also because it minimizes sample loss and allows smaller amounts of proteins to be analyzed. Actually, the ionization of some proteins are suppressed by the presence of other proteins in higher concentrations that suppresses the ionization of proteins of lower abundance, or proteins that suppresses the ionization of glyco- and phosphoproteins etc.

SELDI provides on-chip separation as well as the capability to perform enzymatic reactions directly on the chip. However, there are concerns about the reproducibility of SELDI-TOF mass spectra, especially when normal post processing techniques frequently used with MALDI such as baseline correction are applied. Environmental sources of variation such as humidity can also play a large role.

2.4.3 LIQUID CHROMATOGRAPHY-MS

Liquid chromatography mass spectrometry (LC-MS) is an analytical technique that combines physical separation via liquid chromatography with mass analysis via mass spectrometry. It is the technique of choice for quantitative mass spectrometry-based proteomics because it yields high quality data. LC-MS separates compounds chromatographically before they are introduced to the ion source and mass spectrometer, by means of using liquid mobile phases which ultimately must be volatilized before entering the MS. The mobile phase is liquid, usually a combination of water, organic solvents, and samples instead of gases. The method of coupling high performance liquid chromatography (HPLC) can also be performed with MS. A HPLC simply uses a smaller column that is highly chemically modified to separate on a more precise level than normal LC. Once a sample is injected it goes through a column which separates it based on charge and goes into a drying chamber where the sample is volatilized by a drying gas such as nitrogen. The ions are then collected into a gas capillary where they are collected to be injected further in the system. When the ions proceed out of the gas capillary, the ions go through an area where collision activated dissociation occurs between a skimmer and the capillary, causing the ions to exit individually. The area where the gas capillary and the skimmer meet is the area where volatilization begins. From the capillary, the liquid ions are put through a "Taylor cone". The Taylor cone creates the effect of a fine filament of liquid that volatilizes into a gaseous form by changing its stable liquid droplets to unstable liquid droplets before changing them to gas phase ions. The samples then proceed to an inlet for the mass spectroscopy machine into a quadrupole where they are further separated by charge to mass and then moved to a detector to obtain a mass spectrum. The bottom-up LC-MS approach to proteomics generally involves protease digestion (usually by trypsin enzyme) followed by LC-MS with peptide mass fingerprinting or LC-MS/MS (tandem MS) to derive sequence of individual peptides.

LC-MS-based methods are very powerful and in certain aspects superior or complementary to other approaches such as 2D electrophoresis, for instance. In particular, LC-MS-based methods are capable of capturing both intracellular proteins and membrane proteins and seem to perform especially well for the latter. Since a biological sample can be a mix of thousands of different proteins this feature is crucial for proteomics.

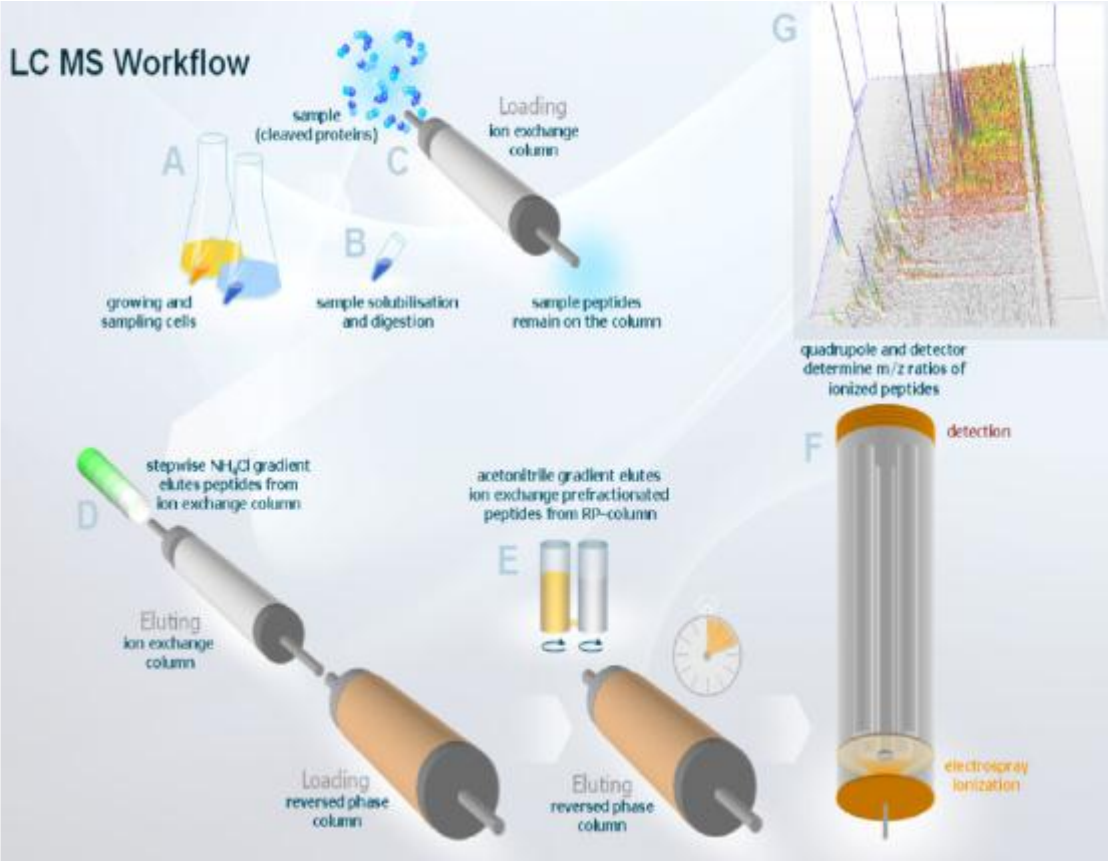


FIGURE 2-21 LC-MS WORKFLOW: (A) GROWTH AND ISOLATION OF THE BIOLOGICAL SAMPLE; (B) PROTEINS IN THE SAMPLE ARE DIGESTED BY PEPTIDASES; (C) SEPARATION OF RESULTING PEPTIDES BY GRADIENT CHROMATOGRAPHY WITH AS A FIRST STEP AN ION EXCHANGE CHROMATOGRAPHY; (D) SECOND SEPARATION STEP IN A REVERSE PHASE COLUMN WITH A GRADIENT APPLIED (E); (F) THE ELUATE ENTERS A QUADRUPOLE AND IN PART REACH THE DETECTOR; (G) DATA VISUALIZATION IN 3D.

The LC-MS workflow is illustrated in Figure 2-21, [22]. In the first step of the processing pipeline, protein molecules are cut into smaller fragments (i.e. peptides), e.g., by the enzyme trypsin. Trypsin cuts at well-defined positions in the amino acid chain (after lysine and also after arginine if not followed by proline), such that the sequences of potential fragments are known when the protein sequence is known. In order to examine the peptides individually, we need to separate them.

Peptide separation is performed by liquid chromatography. A solvent containing the peptides is forced through a separation column (loading). The column contains the stationary phase that binds the peptides. Afterwards, the peptides are washed out of the column by the mobile phase (eluting). The weaker a peptide is bound to the substrate, the faster it gets washed out. Thus, peptides can be separated by their binding properties (e. g. hydrophobicity). The output data of the LC step alone can be displayed using a 2D plot, the chromatogram, where intensity in counts per second is plotted over time.

The masses of the separated peptides can be determined individually using mass spectrometry. MS separates ions by their mass-to-charge ratios. As previously pointed out, In order to analyze peptides by MS, it is necessary to ionize them. Online ionization is realized by ESI. Then, molecules are accelerated and handed to the mass analyzer. The mass analyzer steers the particles to a detector based on their m/z ratio. The detector measures intensity in counts per second. The MS output can be displayed by one 2D plot, the mass spectrum, for each time step. The mass spectrum shows intensity over mass-to-charge ratio (or m/z-ratio).

2.5 TANDEM MASS SPECTROMETRY

In the more modern methods of ionization, like ESI or MALDI, spectra often only contain the ionized molecule with very little fragmentation data and consequently the spectra are of little use for structural characterization of proteins. In these cases, induced fragmentation is required using collision induced dissociation (CID) and tandem mass spectrometry (MS/MS). Fragmentation of gas-phase ions is essential to tandem mass spectrometry and occurs between different stages of mass analysis. There are many methods used to fragment the ions and can result in different types of fragmentation and thus different information about the structure and composition of the molecule. One of the most commonly available tandem mass spectrometers is the triple quadrupole (QQQ) instrument. Here tandem MS is illustrated referring to QQQ, but the general concept is easily extensible to a broader range of mass spectrometers.

Tandem mass spectrometry (MS/MS) involves multiple steps of mass selection or analysis, usually separated by some form of fragmentation. A tandem mass spectrometer is one capable of multiple rounds of mass spectrometry. Multiple stages of m/z separation can be accomplished with individual mass spectrometer elements separated in space or in a single mass spectrometer with the MS steps separated in time.

In tandem mass spectrometry in space, the separation elements are physically separated and distinct, although there is a connection between the elements to maintain high vacuum. These elements can be sectors, transmission quadrupole, or time-of-flight.

In a tandem mass spectrometry in time instrument, the separation is accomplished with ions trapped in the same place, with multiple separation steps taking place over time. A quadrupole ion trap or FTMS instrument can be used for such an analysis. Trapping instruments can perform multiple steps of analysis, which is sometimes referred to as MSn (MS to the n). Often the number of steps, n, is not indicated, but occasionally the value is specified; for example MS3 indicates three stages of separation. You can realize such an experiment using a triple quadrupole instrument.

Tandem mass spectrometry enables a variety of experiments. Although it allows for many uniquely designed experiments some types of experiments are commonly used and built into many commercial mass spectrometers. Examples of these include precursor ion scan, product ion scan and neutral loss scan mode.

2.5.1 PRECURSOR ION SCAN MODE

In precursor ion scan mode, MS2 is used to measure the occurrence of a particular fragment ion (i.e., m/z value) and MS1 is scanning all the m/z values (see Figure 2-22). The resulting spectrum records the ions that are the precursors of the fragments produced in the fragmentation reaction operated by MS2. This experiment is used to detect specific motifs within unknown molecules. In-source fragmentation is often used in addition to tandem mass spectrometry to allow for two steps of fragmentation in a pseudo MS3-type of experiment.

2.5.2 PRODUCT ION SCAN

Post-source fragmentation or product ion analysis is most often what is being used in a tandem mass spectrometry experiment: a mass analyzer can isolate one peptide from many entering a mass spectrometer. It is carried out to analyze only a preselected precursor ion.

The first stage of mass spectrometry (MS1) is set to select a particular m/z and the second stage (MS2) records the mass spectrum of the fragments. The mass spectrum represents the fragments of the ion (or ions) of that particular m/z; this turns the mass spectrometer into an extremely selective detector when used in conjunction with a separation method, such as liquid chromatography mass spectrometry, for example. This offers a much improved sensitivity in comparison with the full MS acquisition.It works almost the same as MRM except here you analyze all the fragment ions.

2.5.3 NEUTRAL LOSS SCAN

In the neutral loss scan both MS1 and MS2 are in operation, but MS2 selects the same m/z as MS1, less the mass of the neutral. The resulting mass spectrum represents all m/z values that lose the neutral by fragmentation.

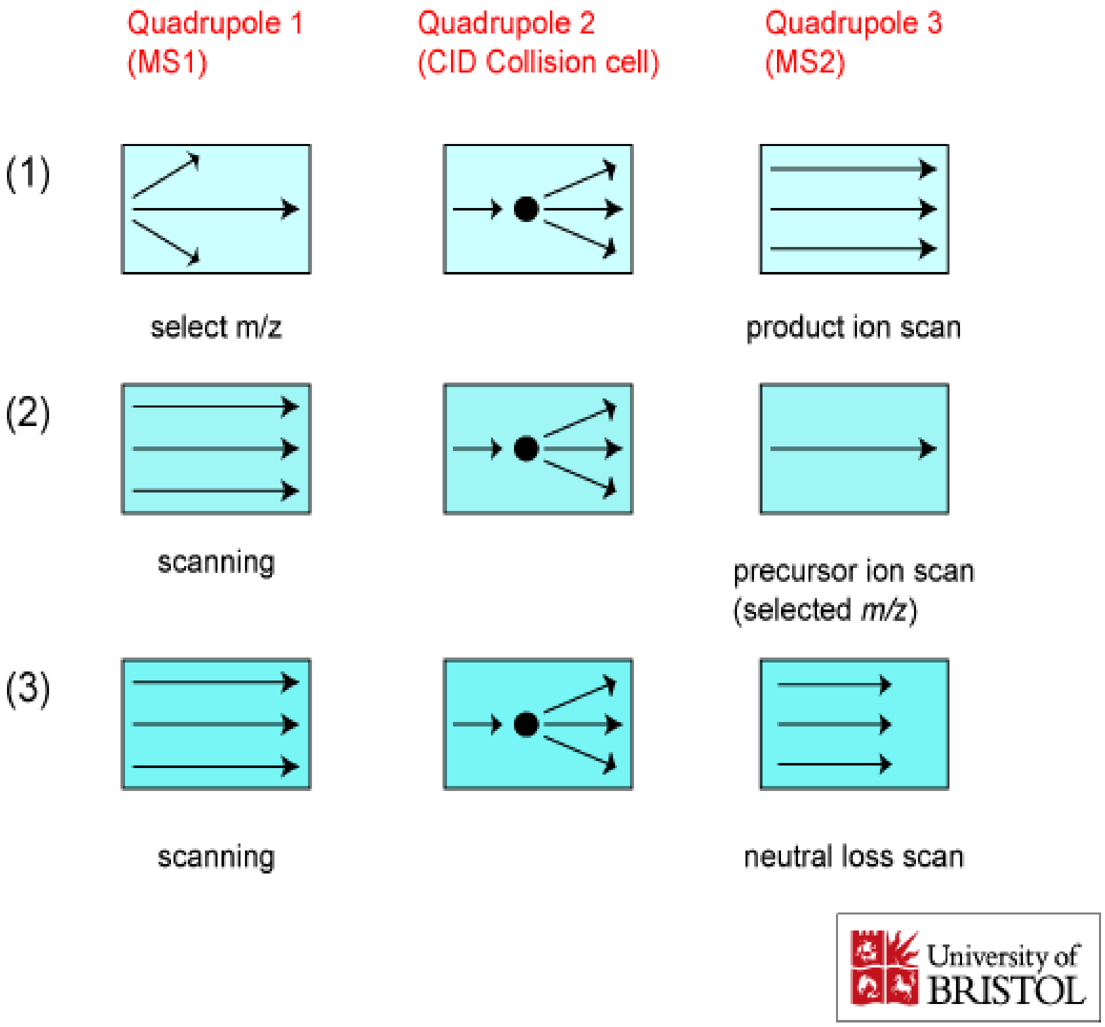


FIGURE 2-22 (1) PRODUCT ION SCAN SCHEMATIC, (2) PRECURSOR ION SCAN SCHEMATIC, (3) NEUTRAL LOSS SCAN.

A peptide sequence tag obtained by tandem mass spectrometry can be used to identify a peptide in a protein database. A notation has been developed for indicating peptide fragments that arise from a tandem mass spectrum (see Figure 2-23). Peptide fragment ions are indicated by a, b, or c if the charge is retained on the N-terminus and by x, y or z if the charge is maintained on the C-terminus. The subscript indicates the number of amino acid residues in the fragment. Superscripts are sometimes used to indicate neutral losses in addition to the backbone fragmentation, for loss of ammonia and for loss of water. Although peptide backbone cleavage is the most useful for sequencing and peptide identification other fragment ions may be observed under certain conditions.

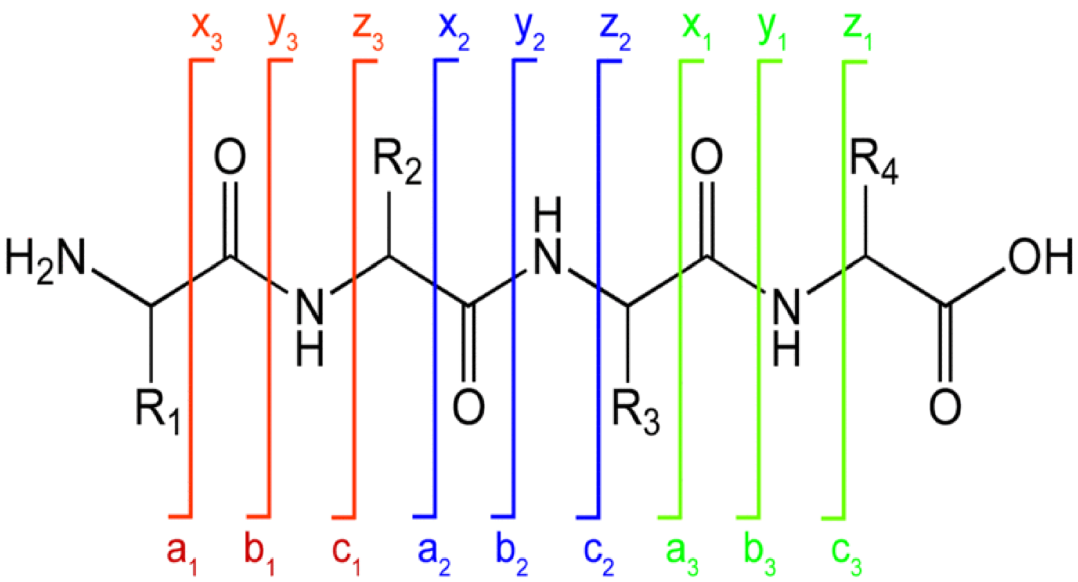


FIGURE 2-23 PEPTIDE FRAGMENTATION NOTATION.

.

3 BIOINFORMATICS FOR QUANTITATIVE MS-BASED PROTEOMICS

During the last decade, it has become available a wide range of technologies which can generate a huge quantity of data potentially able to address relevant questions, e.g., to identify proteins in a biological sample, to quantify their concentration, to monitor post-translational modifications, to measure individual protein turnover, to infer on interactions with other proteins, transcripts, drugs or molecules. Consequently the access, analysis and interpretation of the enormous volumes of MS-based quantitative data are a crucial issue for the advancement of proteomics research.

In this chapter we describe the main bioinformatics topics related to quantitative mass spectrometry-based proteomics data. In particular, we focus on the main quantification strategies to produce quantitative data and on data analysis, especially for profile LC-MS data, which are considered the most appropriate data for quantification aims [23].

Mass spectrometry (MS)-based proteomics plays an ever-increasing role in systems biology, providing information about the qualitative and quantitative content of a biological sample. Since the proteome is involved in functional expression and regulation of systems, MS-based proteomics has become the technique of choice to acquire data to unravel and model biological systems (see Figure 3-1 here below, borrowed from [24]). A major step forward in this direction took place when MS-based proteomics moved ahead from a qualitative approach to a quantitative approach, enabling the association of protein identifications to their quantitative content. Quantitative proteomics is indeed pivotal for many systems biology related fields, such as biomarkers discovery, where researchers aim to recognize differential expression at the proteome and/or genome level: preliminary works suggested that protein abundances are more conserved than transcript abundances (*2*). The cutting edge proteomic technologies will enable researchers to address fundamental biological problems in a systems biology context [1]. In order to properly answer several biological questions, many hypothesis-driven experimental workflows have been designed [25].

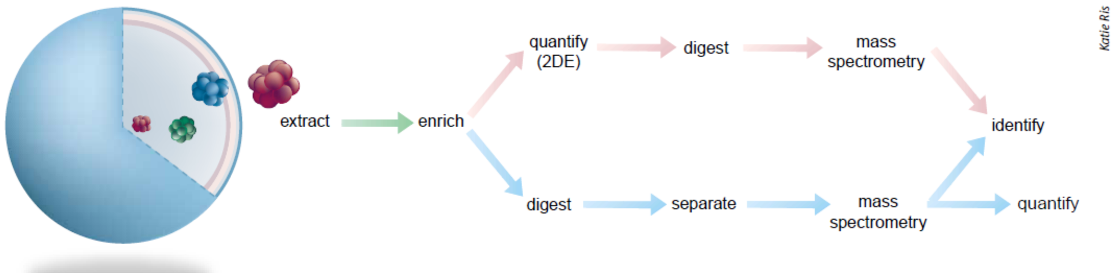


FIGURE 3-1 THE TWO MOST COMMON PROCESSES FOR QUANTITATIVE PROTEOME ANALYSIS ARE SHOWN. IN THE FIRST (TOP), 2DE IS USED TO SEPARATE AND TO QUANTIFY PROTEINS, AND SELECTED PROTEINS ARE THEN ISOLATED AND IDENTIFIED BY MASS SPECTROMETRY. IN THE SECOND (BOTTOM), LC-MS/MS IS USED TO ANALYZE ENZYME DIGESTS OF UNSEPARATED PROTEIN MIXTURES, AND ACCURATE QUANTIFICATION IS ACHIEVED BY LABELING THE PEPTIDES WITH STABLE ISOTOPE.

3.1.1 QUANTIFICATION STRATEGIES

Regarding the quantification strategies, mainly developed to realize biomarkers discovery studies, we could distinguish three approaches, as illustrated in Figure 3-2:

1. the differential stable isotope labeling approach, which analyzes, in the same *Liquid Chromatography-Mass Spectrometry* (LC-MS) run, peptide A and its heavy isotope A*, detected by their characteristic mass difference Δ*m/z* (red) but it heavily depends on the labeling strategy;

2. the spectral counting approach, which computes abundance values counting the number of times a peptide has been identified by tandem mass spectrometry (MS/MS) and compares these across experiments (green) but it is very susceptible to instrument sensitivity;

3. the label-free approach, which extracts peptide signals by tracking corresponding isotopic patterns (along their chromatographic elution profile) across many LC-MS runs (blue) but it has high technical requirements for ensuring reproducibility and perform tracking.

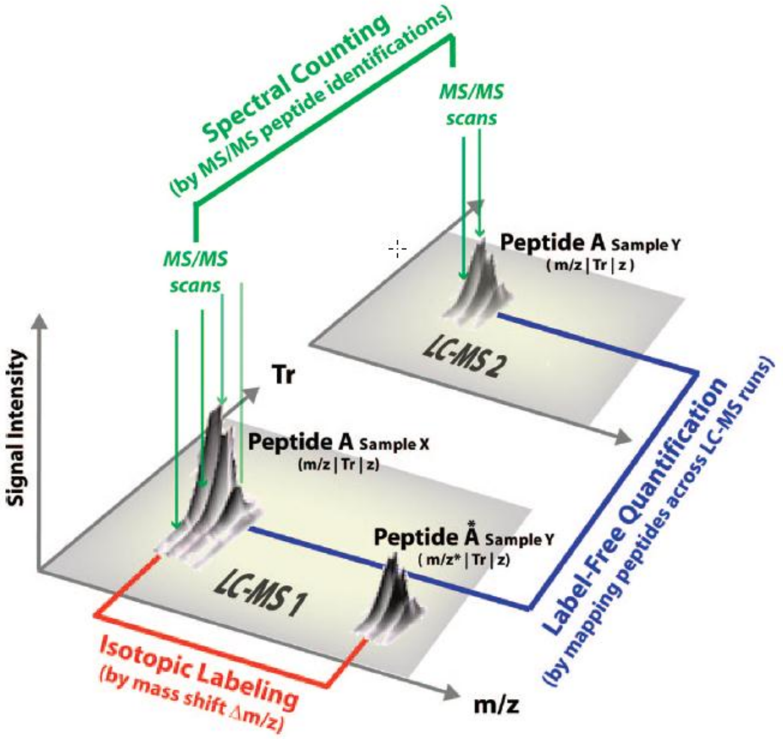


FIGURE 3-2 LC-MS QUANTIFICATION STRATEGIES: THE SPECTRAL COUNTING APPROACH (GREEN); THE DIFFERENTIAL STABLE ISOTOPE LABELING APPROACH (RED); THE LABEL-FREE APPROACH (BLUE).

3.1.1.1 STABLE ISOTOPE AND ISOBARIC LABELING

Quantitative proteomics allows for the determination of both the identity and relative quantity of particular components across different samples. Stable isotopes labeling [24] is ideal for use in quantitative proteomics because “light” and “heavy” isotopes have the same chemical behavior and properties and their mass shift is easily detectable by the mass spectrometer. Since they are chemical identical, those labeled peptides coelute, thus samples can be merged labeled with “light” and “heavy” isotope tags and to process them in a single run. The comparison (i.e., ratio) of the relative intensities of the “heavy” versus “light” labeled peptides in the MS signal provides the quantification. Isotopic labeling strategies enable the highly accurate quantification of LC-MS experiments since analysis is performed on single LC-MS runs where peptide pairs can be very accurately detected by distinct mass shifts characteristic to the utilized label (see Figure 3-3).

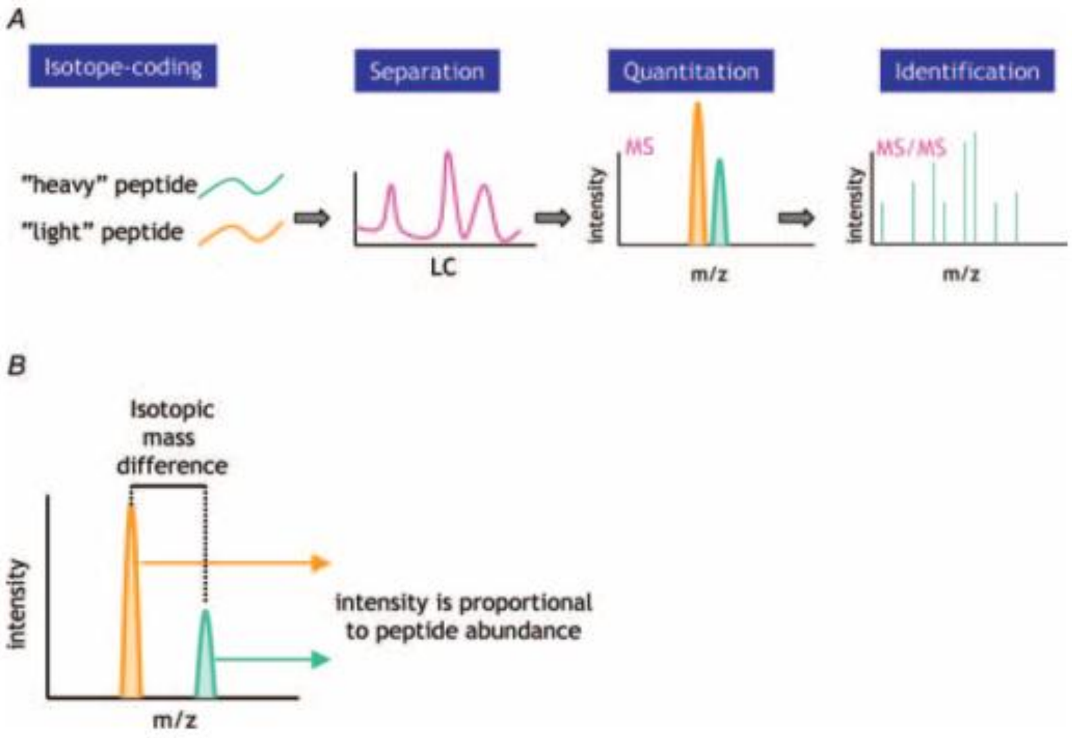


FIGURE 3-3 ISOTOPIC LABELING IN QUANTITATIVE PROTEOMICS (A); THE MASS SHIFT HELPS TO DISTINGUISH THE SIGNALS BELONGING TO THE DIFFERENT ISOTOPE TAGS (B).

STABLE ISOTOPE-LABELED AMINO ACIDS IN CELL CULTURE

SILAC (stable isotope-labeled amino acids in cell culture) [26] involves a metabolic incorporation of isotopically heavy amino acids into proteins Figure 3-4. In SILAC labeling two populations of cells are grown in the same type of culture medium, except that in one set, one or more essential amino acids are replaced by a version containing heavy atoms (e.g. ^{13}C): for this reason is considered an "in vivo" kind of labeling. Specifically, cell cultures are grown in media containing either light ^{12}C or heavy ^{13}C labeled arginine and lysine to metabolically incorporate the modified amino acids into proteins through the metabolic cycle. The generated isotopic peptide pairs are then detected by mass shifts of multitudes of 6 mass units. Since the label is added at a very early stage of the experiment, this technology circumvents the introduction of additional error sources through extra experimental sample processing steps. However, SILAC labeling is largely limited to biological material that can be grown in culture and thus is not generally applicable to tissues, body fluids, or clinical applications. Recently, metabolic conversion of the stable isotope labeled peptide has also been reported, resulting in the added label in unexpected amino acids [27].

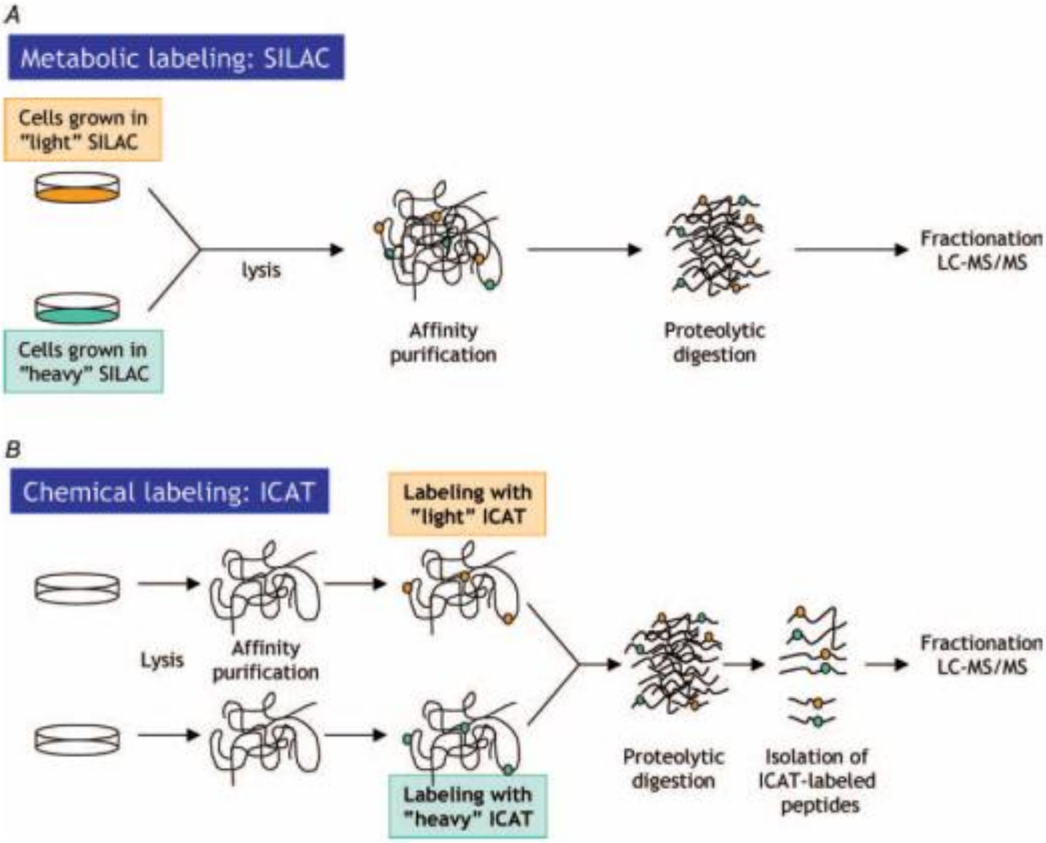

A
Metabolic labeling: SILAC
Cells grown in "light" SILAC
Cells grown in "heavy" SILAC
lysis
Affinity purification
Proteolytic digestion
Fractionation LC-MS/MS
B
Chemical labeling: ICAT
Lysis
Affinity purification
Labeling with "light" ICAT
Labeling with "heavy" ICAT
Proteolytic digestion
Isolation of ICAT-labeled peptides
Fractionation LC-MS/MS

FIGURE 3-4 SILAC AND ICAT SCHEMATICS.

ISOTOPE CODED AFFINITY TAGGING

Another common technique, which instead is “in vitro”, is ICAT (isotope coded affinity tagging) [24] which involves a chemical attachment of isotopic tags to proteins or peptides in solution Figure 3-4. The nature of ICAT tag may vary a lot, but the reagents are generally composed of a reactive group used to covalently attach the tag to peptides, a linker group containing the isotope, and an affinity handle such as biotin (see Figure 3-5, borrowed from [28]). All different reagents specifically target cysteine groups. The labeled peptides differ in their molecular weight by 8 mass units, and in newer versions by 9 mass units. More recently, a number of variants of this concept have been developed in which sets of reagents differ in specificity, structure, mass difference, and number of isotopic forms.

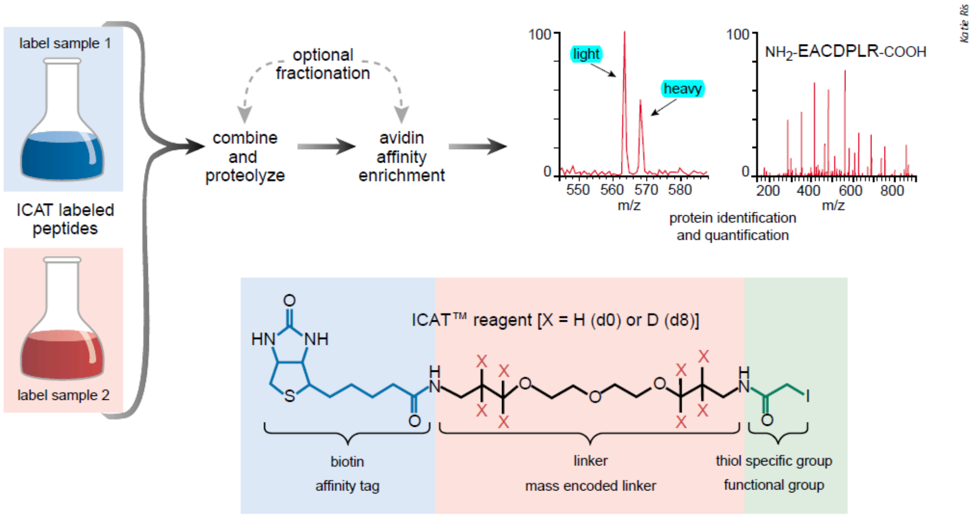


FIGURE 3-5 SCHEMATIC REPRESENTATION OF THE ICAT LABELLING STRATEGY.

ISOTOPE CODED PROTEIN LABELS

In the isotope coded protein labels (ICPL) strategy, two protein mixtures obtained from two distinct cell states or tissues are first individually reduced and alkylated to denature the proteins and to ensure easier access to free amino groups that are subsequently derivatised with the deuterium free (light) or deuterium containing (heavy) form, respectively, of the ICPL reagent (se Figure 3-6). After combining both mixtures, any separation method can be adopted to reduce the complexity of the sample on the protein level and, after digestion, on the peptide level followed by high throughput MS/MS.

The ICPL strategy is based on stable isotope labeling of free amino groups in intact proteins and has the capability to become the basis for comprehensive high-throughput proteome analysis for several reasons. First, employing Nic- NHS (nicotinoyloxy succinimide) as a labeling reagent enhances MS sensitivity, making this tag ideally suited for the analysis of low abundant proteins or when the amount of the sample is limited. Second, the ICPL strategy enables multiplexed analysis of three samples in one single assay for increased throughput. Third, the number of lysine residues of labeled peptides, that can be easily calculated from the mass gap of an isotope peptide pair, serves as a strong constraint in database searches. Fourth, since ICPL is based on stable isotopic labeling of intact proteins at a very early stage, there are essentially no limitations in terms of compatibility with separation and analyzing techniques or protein samples to handle. Other protein isotope labeling approaches described to date, in particular the ICAT strategy, have also been shown to correctly quantify the abundance of proteins in complex mixtures. However, the main limitation of these techniques results from their specificity for the rare sulfhydryl groups in proteins. As a consequence, these approaches fail to quantify a considerable number of proteins that contain no or only a few cysteine residues. Conversely, the ICPL method has the potential to quantify almost every protein. ICPL approach is based on isotopic labeling of intact proteins and is accordingly compatible to all protein or peptide separation techniques currently employed in proteome research.

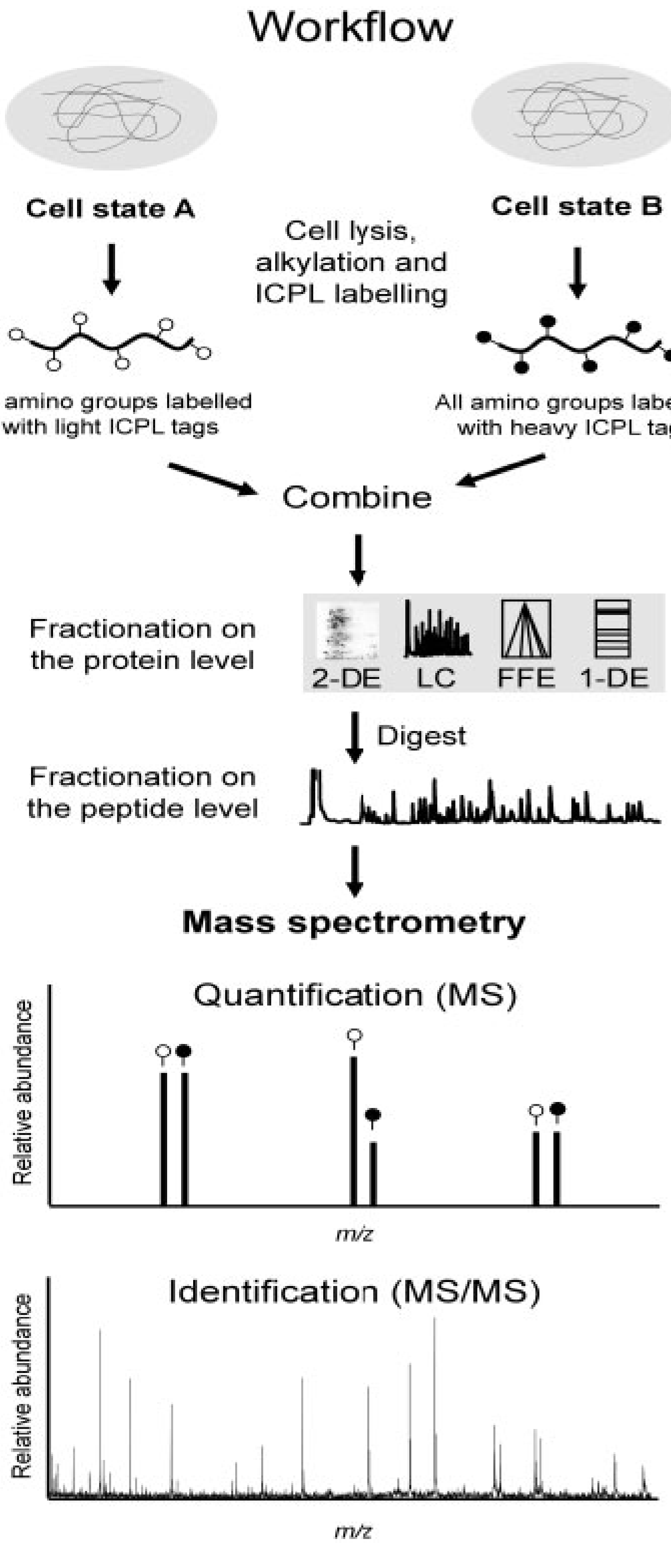

Workflow
Cell state A
Cell state B
Cell lysis, alkylation and ICPL labelling
All amino groups labelled with light ICPL tags
All amino groups labelled with heavy ICPL tags
Combine
Fractionation on the protein level
2-DE
LC
FFE
1-DE
Digest
Fractionation on the peptide level
Mass spectrometry
Quantification (MS)
Relative abundance
m/z
Identification (MS/MS)
Relative abundance
m/z

FIGURE 3-6 OVERVIEW OF THE ICPL WORKFLOW. ANY SEPARATION METHOD CAN BE EMPLOYED TO REDUCE COMPLEXITY ON THE PROTEIN LEVEL (E.G. 1-DE OR 2-DE, FREE FLOW ELECTROPHORESIS (FFE), LC) AND, AFTER PROTEOLYSIS, ON THE PEPTIDE LEVEL (E.G. MULTIDIMENSIONAL LC) FOLLOWED BY MS/ MS.

The 8-plex iTRAQ labeling allows the simultaneous quantification of eight biological samples [29]. The isobaric (i.e., having the same mass) reagent reacts with primary amino groups and produces in the MS/MS fragmentation spectrum eight different unique reporter groups, one per reagent flavor, at 113, 114, 115, 116, 117, 118, 119, and 121 *m/z*. iTRAQ labeling does not increase the sample complexity because the reagent is based, and relies, compared to isotopic labeling, on a fully MS/MS-dependent workflow. Therefore, after mixing, in MS^1, the peptides appear as a single precursor. However, when fragmented during MS^2, in addition to the normal fragment ions, the reporter regions dissociate to produce ion signals which provide quantitative information regarding the relative amount of the peptide in the samples. Thus, only peptides are quantified that were subjected to CID fragmentation and could be successfully assigned to a peptide sequence. In the figure below the scheme of isobaric labeling is illustrated and compared to the one of isotopic labeling. Once the reporter ions are recognized and the ratio computed, downstream analysis is the same as for isotopic labeling, as shown in Figure 3-7.

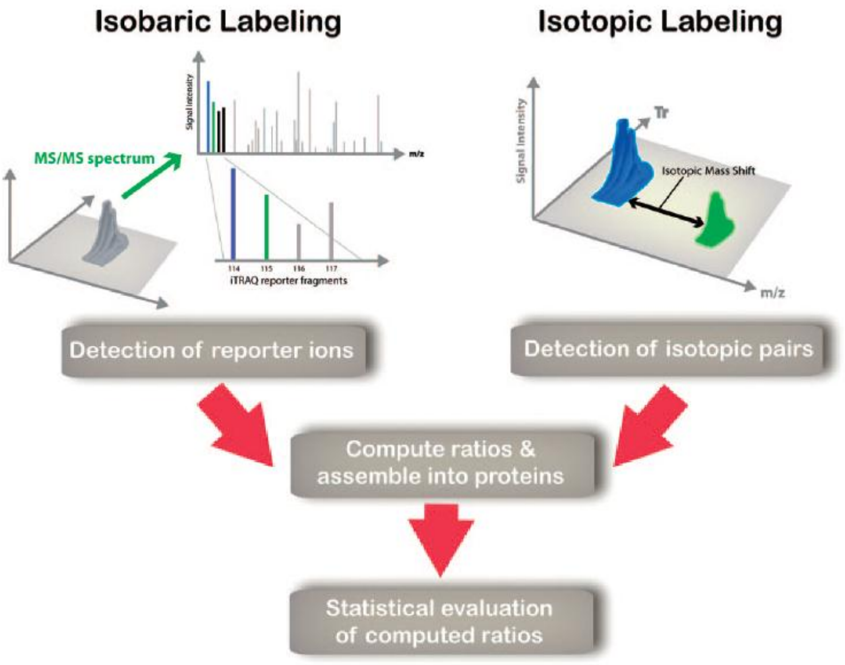


FIGURE 3-7 THE QUANTIFICATION PRINCIPLES OF ISOBARIC AND ISOTOPIC LABELING ARE SCHEMATICALLY ILLUSTRATED. ISOBARIC LABELING GENERATES IN THE MS/MS SPECTRA DIFFERENT REPORTER IONS THAT ARE USED TO CALCULATE PEPTIDE ABUN- DANCE VALUES BETWEEN DIFFERENT SAMPLES. ISOTOPIC APPROACHES DIFFERENTIALLY LABEL PEPTIDES OR PROTEINS FROM TWO SAMPLES (GREEN/ BLUE) TO PRODUCE ISOTOPIC PAIRS OF CHARACTERISTIC MASS SHIFTS.

3.1.1.2 SPECTRAL COUNTING

The concept of semi quantitative analysis was introduced for shotgun proteomics, in which the instrument control system of the mass spectrometer autonomously selects a subset of peptide precursor ions detected in a survey scan (MS1 scan) for collision induced fragmentation (CID) following predetermined rules (typically, the 1–5 most intense precursor ions) [30]. This quantification strategy is based on the hypothesis that the MS/MS sampling rate of a particular peptide, i.e., the number of times a peptide precursor ion is selected for CID in a large data set, is directly related to the abundance of a peptide represented by its precursor ion in the sample mixture [31]. This approach, also termed spectral counting, therefore transforms the frequency by which a peptide is identified into a measure for peptide abundance. Spectral counts of peptides associated with a protein are then averaged into a protein abundance index. Spectral counting approaches have most frequently been used for the analysis of low to moderate mass resolution LC-MS data and serve therefore as a convenient, fast, and intuitive quantification strategy. A critical point in spectral counting is how spectral counts are computed if only a small number of peptide identifications are available. It especially holds for the quantification of low-abundance proteins since the selection of precursor masses for MS/MS analysis in shotgun experiments is skewed toward peptides of high abundance and the identification of low-abundant peptides is very irreproducible between LC-MS experiments. Corresponding abundance indexes of such low-abundant proteins are therefore unreliable since they are obtained from spectral counts of only a small number of peptide identifications.

3.1.1.3 LABEL-FREE

With the evolution of mass spectrometers toward high mass precision instruments, label-free quantification of LC-MS data has become a very appealing approach for the quantitative analysis of biological samples. Typically, peptide signals are detected at the MS1 level and distinguished from chemical noise through their characteristic isotopic pattern. These patterns are then tracked across the retention time dimension and used to reconstruct a chromatographic elution profile of the monoisotopic peptide mass. The total ion current of the peptide signal is then integrated and used as a quantitative measurement of the original peptide concentration. In principle, every peptide signal within the sensitivity range of the MS analyzer can be extracted and incorporated into the quantification process independent of MS/MS acquisition [32]. This leads to an increased dynamic range of the peptide detection and largely reduces the undersampling problem common to the previously described MS/MS-based approaches. Label-free strategies were in most cases applied to data acquired on mass spectrometers equipped with the new generation of time-of flight, Fourier transform-ion cyclotron resonance, or Orbitrap mass analyzers. Measurements on these MS platforms reach very high resolution power and mass precision.

In contrast to differential labeling, every biological specimen needs to be measured separately in a label-free experiment. The extracted peptide signals are then mapped across few or multiple LC-MS measurements using their coordinates on the mass to charge and retention time dimension. The efficiency of the peptide tracking depends on the available mass resolution of the utilized mass spectrometer. Data from high mass precision instruments greatly facilitate this process and increase the certainty of matching correct peptide signals across runs. In addition to the *m/z* dimension, the retention time coordinate is used to map corresponding peptides between runs. Therefore, the consistency of the retention time values over different LC-MS runs is a crucial factor and has led to the development of various alignment methods to correct chromatographic fluctuations [32]. Finally, sophisticated normalization methods are important to removing systematic artifacts in the peptide intensity values between LC-MS measurements.

3.1.2 MS SETUP FOR QUANTIFICATION

LC-MS-based techniques are commonly used for quantitative analyses. Indeed, LC-MS has paved the way to quantify a large number of peptide elements of biological samples in an automated and high-throughput mode. Here below the main 3 setups for running the MS experiment are briefly illustrated. In particular, tandem MS arrangements are different from an approach to the other and are chosen in relation to the aim of the experiment itself.

3.1.2.1 FULL MS SCAN MONITORING

Proteomic studies are commonly performed using a shotgun approach, in which the sample proteins are enzymatically degraded to peptides, which are then analyzed by mass spectrometry (MS). Data are collected in full MS scan mode. Thereby, a subset of the peptides present in the sample is automatically and in part stochastically selected by the mass spectrometer in a process referred to as data-dependent precursor selection. The simplest method to quantify analytes by

LC-MS is the use of eXtracted Ion Chromatograms (XIC). Data are processed post-acquisition, to reconstruct the elution profile for the ion(s) of interest, with a given m/z value and a tolerance. XIC peak heights or peak areas are used to determine the analyte abundance.

3.1.2.2 SELECTED ION MONITORING

Selected ion monitoring is performed on scanning mass spectrometers, by restricting the acquisition mass range around the m/z value of the ion(s) of interest. The narrower the mass range, the more specific the SIM assay. SIM experiments are more sensitive than XICs from full scans because the MS is allowed to dwell for a longer time over a small mass range of interest. Several ions within a given m/z range can be observed without any discrimination and cumulatively quantified; quantification is still performed using ion chromatograms.

3.1.2.3 SELECTED REACTION MONITORING

Selected reaction monitoring (SRM), also called multiple reaction monitoring, is emerging as a technology that ideally complements the discovery capabilities of shotgun strategies by its unique potential for reliable quantification of analytes of low abundance in complex mixtures [33-36]. In an SRM experiment, a predefined precursor ion and one of its fragments are selected by the two mass filters of a triple quadrupole instrument and monitored over time for precise quantification.

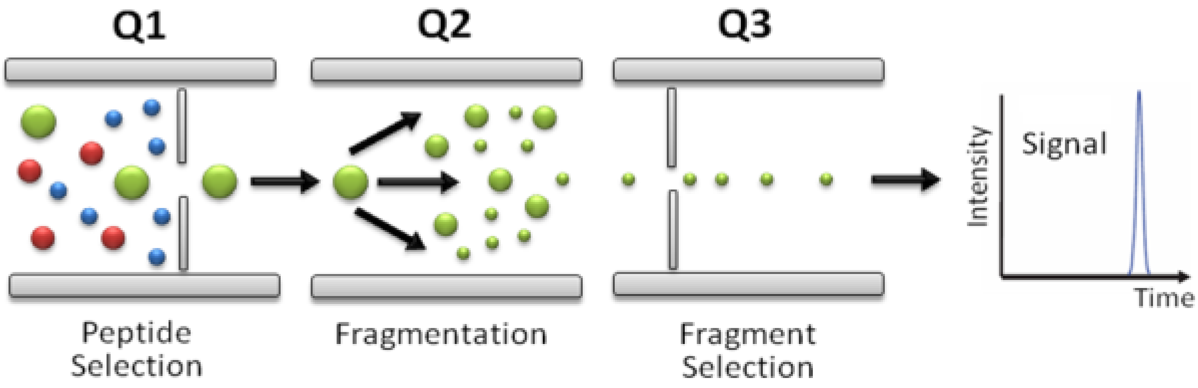


FIGURE 3-8 SRM SCHEMATIC. TWO MASS ANALYZERS ARE USED AS STATIC MASS FILTERS, TO MONITOR A PARTICULAR FRAGMENT ION OF A SELECTED PRECURSOR ION, WHEREAS THE SECOND MASS ANALYZER IS USED AS A COLLISION CELL.

SRM exploits the unique capabilities of triple quadrupole MS for quantitative analysis. In SRM, the first and the third quadrupoles act as filters to specifically select predefined m/z values corresponding to the peptide ion and a specific fragment ion of the peptide, whereas the second quadrupole serves as collision cell (see the schematic in Figure 3-8). Several such transitions (precursor/fragment ion pairs) are monitored over time, yielding a set of chromatographic traces with the retention time and signal intensity for a specific transition as coordinates. The two levels of mass selection with narrow mass windows result in a high selectivity, as co-eluting background ions are filtered out very effectively. Unlike in other MS-based proteomic techniques, no full mass spectra are recorded in triple quadrupole-based SRM analysis. The non-scanning nature of this mode of operation translates into an increased sensitivity by one or two orders of magnitude compared with conventional ‘full scan' techniques. In addition, it results in a linear response over a wide dynamic range up to five orders of magnitude. This enables the detection of low-abundance proteins in highly complex mixtures, which is crucial for systematic quantitative studies. Proteins of interest can be detected with a much increased sensitivity and at a higher throughput than other techniques.

A figure from [37] summarizing several quantification methods is reported in Figure 3-9. Some of them weren’t explained here, since only the most common strategies were illustrated.

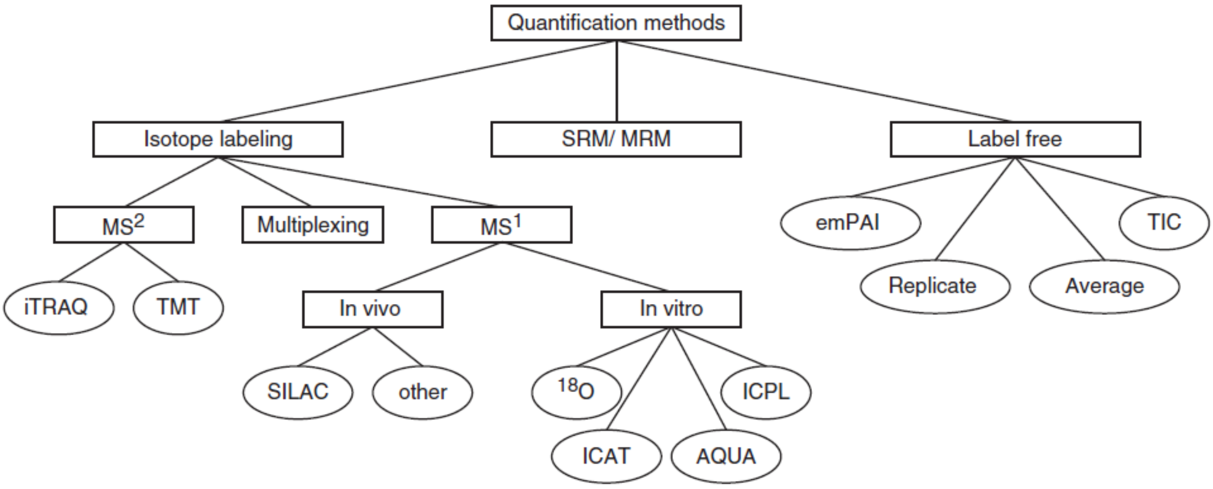


FIGURE 3-9 A FIGURE SUMMARIZING SEVERAL QUANTIFICATION STRATEGIES: SORTED ACCORDING TO THE PRESENCE OF LABEL, THEN WHERE THE LABEL DISCRIMINATES THE PEPTIDES (MS LEVEL) AND FINALLY WHERE THE LABEL IS APPLIED.

3.2 DATA ANALYSIS

In this paragraph we aim to provide an overview of the data processing involved in the quantitative MS-based proteomics field, highlighting the potential pitfalls, strengths and sensitive points. Indeed, with the availability of mass spectrometry methods to analyze complex biological samples at a large-scale level a necessity arose for computational methods to analyze and statistically evaluate data generated from LC-MS experiments, thus catalyzing a new research direction in the field of bioinformatics.

3.2.1 LC-MS DATA

Profile LC-MS datasets are considered the most suitable MS-based data for quantification aims (8). In this paragraph we focus on their main features: the 3D structure and the profile acquisition mode.

3.2.1.1 3D STRUCTURE

The output data of the LC step alone can be displayed using a 2D plot, the chromatogram, where intensity in counts per second is plotted over time. MS output can be displayed by one 2D plot, the mass spectrum, for each time step. This 2D plot shows intensity (i.e., counts per second measured by the detector) over mass-to-charge ratio (or m/z-ratio). Data coming out of a liquid chromatograph is a function over time. When delivering the data from the detector to a computer system, the values are given at discrete points in time that are not distributed equidistantly. The number of points can be in the range of many thousands. The data coming out of the mass spectrometer is a function over the m/z-ratio. The intensity is measured in counts per time and stored as an intensity list for discrete m/z-ratios: the m/z-ratios are not distributed equidistantly, either. Their number depends on the experimental setup and can be in the range of tens to hundreds of thousands. Instead of generating two-dimensional graphs for each point in time, we use a three-dimensional setup, where the intensity is shown as a height field over the dimensions time and m/z-ratio.

Therefore, LC-MS provides intensity data on a 2D (t, m/z) domain, since LC separates proteins along retention time dimension (temporal index) based on their chemical-physical properties, while MS separates proteins based on their mass over charge (*m/z* index) ratios. Unfortunately, the m/z-ratios vary from one point in time t_i to the subsequent point in time t_{i+1}, and even the number of values in the m/z-ratio dimension varies significantly. Thus, when looking at a two-dimensional domain with the dimensions being m/z ratio and time, data positions are scattered in one dimension and non-equidistant in the other dimension.

Generally, all kind of MS experiments have a "temporal" index related to the experimental time at which the MS acquisition takes place, even if the LC separation is not coupled to the MS. Thus, we can conceptually view an LC-MS (or, more generally, MS) dataset as a matrix, where the rows are indexed by retention times (scan if MS), the columns by m/z values, and the entries are intensity values (see Figure 3-10). A generic entry is denoted as (*rt*, *mz*; *I*), where *rt* and *mz* are the row and column indices, and *I* is the intensity value.

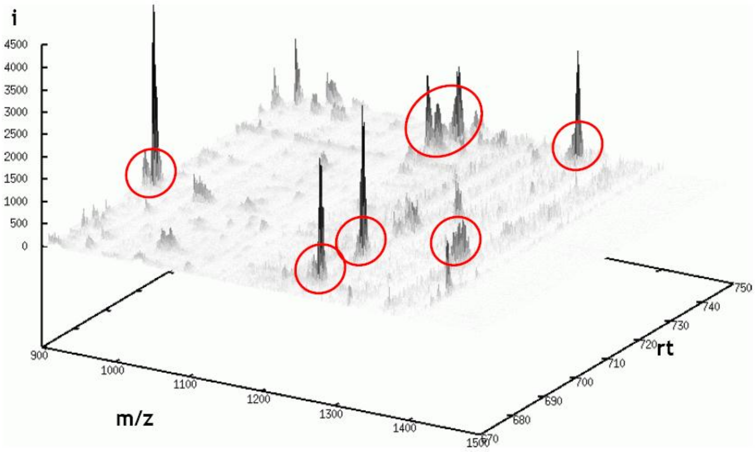


FIGURE 3-10 A 3D REPRESENTATION OF AN LC-MS MAP: RED CIRCLES ARE PICKING THE DATA FEATURES REFERRING TO PEPTIDE DISTRIBUTIONS.

3.2.1.2 PROFILE VS CENTROID ACQUISITION MODE

The profile acquisition mode is an almost continuous acquirement of the observed ion current. Profile mode data is the most informative way to save and store the data, it is the raw signal acquired during the mass spectrometric experiment itself. Many researchers regard raw/profile data as the only data source rich enough to perform a meaningful quantitative analysis [23].

Rather than retaining abundance information, peaks are frequently centroided. Centroid mode mass spectrum is created when the mass spectrometer software is instructed to automatically find the centroid of all peaks as they are recorded. It is frequently used to reduce the size of an

LC-MS dataset as it is recorded. Centroid mass spectra contain discrete peaks of zero width and most of their original informative content is permanently lost (see Figure 3-11). Such a step is suboptimal for downstream data analysis: underlying methodologies of machine learning and statistical techniques are intended to account for random variation caused by noise, and their performance is likely deteriorated by using them with centroided MS data. Incorporating richer information would likely improve analytical performance, albeit at the cost of more computation.

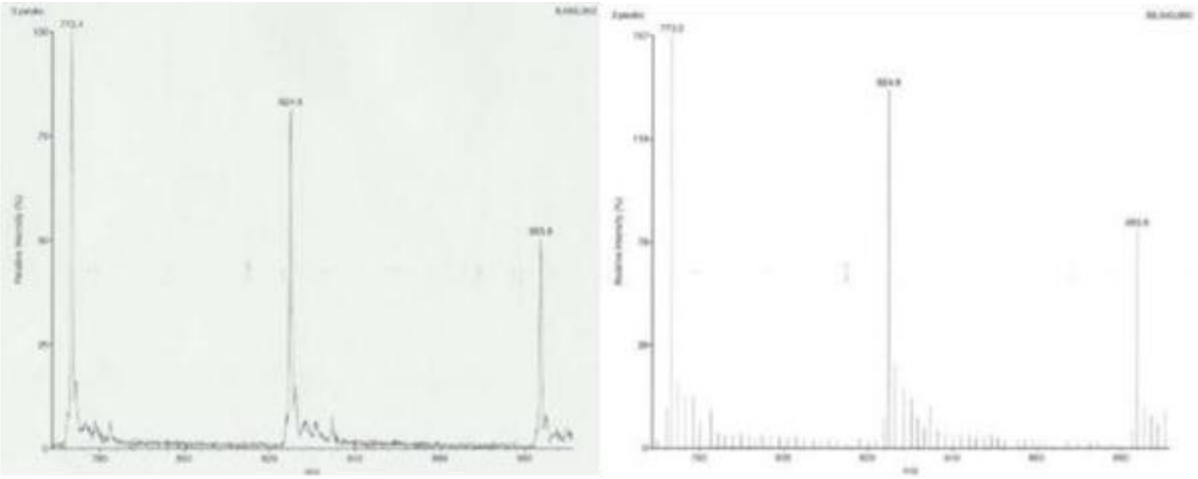


FIGURE 3-11 A MASS SPECTRUM REPRESENTED IN PROFILE MODE (LEFT SIDE) AND THE CORRESPONDING CENTROID MODE DATA (RIGHT SIDE).

3.2.2 LC-MS DATA ANALYSIS

Regarding signal processing we can recognize three major steps to understand how a data analysis start-to-finish approach should be designed:

1. low-level processing involves raw LC-MS signal with some basic pre-processing, such as m/z quantization, filtering, formation of a data matrix and background subtraction to minimize noise;

2. mid-level processing steps, such as data normalization, alignment in time, peak detection, peak quantification, peak matching, and error models, to facilitate profile comparisons;

3. high-level processing is applied to data that has been fully massaged for use in conjunction with machine-learning techniques (for sample classification and biomarker discovery) or more

traditional statistical techniques such as significance testing of individual features (e.g., peptide abundance), multiple testing, and choice of feature space.

The substantial collection of methods developed for processing non chromatographic MS data (e.g. MALDI and SELDI studies) is in many cases transferable to LC-MS data, which are commonly viewed as a time series of static MS spectra. Most of the low- and mid-level processing methods reported to date, however, have been performed parenthetically as a means to the larger goal of sample classification or biomarker discovery and hence have not been rigorously studied.

Here we describes only aspects related to low level processing, which is necessary as a preliminary step to perform the quantitative data analysis considered in the following of this thesis.

3.2.2.1 LOW-LEVEL PROCESSING

Given LC-MS data structure it appears natural to convert LC-MS datasets into a matrix format, with columns representing m/z values and rows representing time (or scan). This matrix formation often involves binning nominal m/z values, because retaining all m/z possible values would lead to a huge, sparsely populated, matrix, while time values can normally be left unchanged because these are usually not too numerous and because many m/z values typically correspond to a given time point.

BINNING: QUANTIZATION OF M/Z VALUES

An optimal bin width would be large enough to keep the matrix tractable and not too sparse, but small enough that individual m/z values remain informative (i.e., not collapsing information too much). Such trade-off depends on the MS instrumentation used to acquire data. As a binning strategy it is possible to opt for evenly spaced bins in either native or log *m/z* space.

Binned intensity can be computed taking the mean or the summation of the intensity values in the bin. Binning can be seen as a sort of sub-sampling, or a low-pass filtering process, which removes part of the spiky nature of the MS data.

No methods have been reported for evaluating optimal bin width, nor for determining the sensitivity of further calculations to this parameter. The choice of the “bin width” plays a major role in the subsequent analysis, since it can heavily shrink the original data dimensionality. There is no rule which gives an a priori information on which choice will yield the best results, and while a large value for the “bin width” helps in reducing the original signal size, it may also lead to the loss of relevant peaks. If possible, it is therefore better to keep data in their raw form, which is the most informative.

BACKGROUND SUBTRACTION AND SIGNAL SMOOTHING

Given that LC-MS is subject to background chemical and electronic noise, together with systemic contaminants in the LC mobile phase (column solvent), methods for noise reduction and signal enhancement are commonly used. Fortunately signal filtering is a mature field from which a variety of techniques are applicable. The theory of digital signal processing is based on the assumption that data were sampled at regular time intervals, which is not necessarily the case for many LC-MS experiments, even if there are experimental efforts to obtain a uniform sampling rate. Filtering may nevertheless be useful, provided extra care is taken to account for uneven sampling rate.

Conceptually, signal filtering and baseline subtraction can be performed in both time (scan number) and m/z dimension. At the state of art two are the approaches mostly applied: either subtracting a fitted, additive baseline model or applying digital filters to smooth and enhance the MS signal.

Various filters for data smoothing along the LC time axis have been implemented including: simple “moving average,” median, moving geometric mean filters, loess smoother (a moving window filter with tricubic kernel of weights), or the Savitzky-Golay filter, which preserves high-frequency content, like peaks, by fitting a high-order polynomial to the data over a local window [38].

Manual delineation of background is a subjective, tedious, and error-prone process, and inconsistent with high-throughput analysis. Moreover even if the Savitzky-Golay filter is widely used, it would be necessary the establishment of a robust, but computationally fast, statistically based method to set its parameters trying to avoid over fitting.

Since filtering efforts published to date were performed only on one data dimension, it will be interesting to see if filtering independently in both axes (time and *m/z*) or simultaneously with a 3D approach is more beneficial [32].

3.2.2.2 MID-LEVEL PROCESSING

Extracting peaks from LC-MS signal both reduces the dimensionality of the data, which can simplify downstream analysis, and gives intuitive meaning to data features: it is advantageous to detect and quantify two-dimensional peptide peaks in LC-MS signal for use as input to classification algorithms, biomarker discovery, or global proteomic comparisons using a unified reporting schema.

PEAK DETECTION

Peak detection has generally been performed in a rather ad hoc manner, with little evaluation of the efficiency of the diverse methods or parameter choices. The algorithms employed to date make no use of a priori or learned information with regards to peak shape, along either the time or m/z dimensions, and in some cases ion intensity values are only exploited very indirectly. Rather than retaining abundance information, peaks are frequently binarized, because it helps to overcome noise in the signal. Such a step is lossy and is likely suboptimal for downstream analysis: underlying methodologies of machine learning and statistical techniques are intended to account for random variation caused by noise, and their performance is likely deteriorated by using them with binarized MS data. Incorporating richer information would likely improve analytical performance, albeit at the cost of more computation.

Peak detection, followed by quantification, even if done optimally, does not guarantee linearity of peak signal relative to analyte concentration due to possible ion suppression effects. Different compounds have differential ionization capabilities and therefore intensity of your ion is not a direct correlation to concentration. Nevertheless, compelling evidence of linearity of extracted LC-MS peak intensities, at least for spiked reference proteins, has been established using certain data processing methods and technological platforms [39-41] .

In such an undefined scenario, groups of researchers made their proposal, creating very different methods to handle somehow the peak detection and quantification issues: here below there are some fascinating examples.

Radulovic et al. used an iterative coarse-to-fine strategy to extract two dimensional (in time and m/z) peaks from LC-MS data [41]. Neighbouring points in the data matrix deemed to be signal were combined to form peaks at the roughest level, and then iteratively through each of the more refined levels, with a bisection method used to avoid spurious peak mergers.

Wang et al. detected LC-MS peaks based on coinciding local signal maxima, in time and m/z; local maxima are defined as an increase in ion abundance greater than a pre-specified threshold over a predefined range [40]. Peaks were then quantified either by summing intensity over the component elution time or based on the maximum peak height.

Leptos et al. in the Mapquant [42] segmented the 2-D map obtained visualizing the LC-MS data using the watershed segmentation algorithm [43]. The function implementing this algorithm returns a 2-D labelled non-gray-scale map that has the form of a mosaic, which, along with the noise-filtered 2-D-map from the previous step and information about the rectangular circumscribed boundaries of the segment, can be used to cut out a so-called segment map. Peaks that are well resolved are confined into individual segment maps whereas overlapping peaks are confined into common segments. The latter is possible through a morphological opening operation of the noise-filtered 2-D-map prior to segmentation. The peak detection algorithm uses concepts from mathematical morphology such as the structuring element theory.

PEAK MATCHING

Peak matching is another related topic relevant to quantitative proteomic comparisons. Reliable peak matching is crucial for label-free approaches to quantification. To measure reproducibility of peptide signal, experimental peaks must be matched across LC-MS datasets.

Naive methods, based on simple proximity (in time or m/z dimension), are reported to be effective [40].

For instance, Radulovic et al. used MS/MS-derived sequence identities to verify the correct matching of ~200 putative peptides across multiple samples [41]. However, given that MS/MS targets prominent peaks, this assessment is likely biased. Incorporation of prior knowledge of peak shape, instrument m/z drift, and a more-probabilistic formulation might significantly improve the effectiveness of peak detection, quantification, and matching.

A good alignment among datasets would help a better and easier peak matching.

DATASETS ALIGNMENT

A challenging task is to compare multiple LC-MS profiles matching corresponding peptide features (i.e., peak matching) from different experiments, that, for example, can be used to identify discriminating peptides between distinct biological groups or to quantify peptides in label-free approaches. Because the sequence identifications of the peptide are often unavailable at this stage, one relies on RT and m/z to match corresponding peptides across different samples. However, the retention time of a specific peptide depends on instrument conditions as well as the underlying composition of the mixture; variation in RT between experiments is often non-negligible even when all samples are processed by the same LC-MS system. LC fractionation is inherently variable since considerable dispersion could affect peptide retention times. Elution patterns can become locally compressed and/or expanded in complex, nonlinear ways by differences in chromatography performance due to changes in ambient pressure and temperature. Even under ideal conditions, MS duty cycles are finite and sampling is not necessarily constant, resulting in spectral capture at different time points across an eluting peak even between repeat analyses. This variation can affect peak discrimination and global proteomic comparisons. Thus, to maximize the benefits of LC-MS, one needs to deal with the inherent variability in the time axis (i.e., recorded retention time or scan headers). To a lesser extent, m/z of a peptide also varies as a result of instrument noise, although this is far less of a problem than variations in time. For these reasons, a prerequisite for quantitative analysis of multiple LC-MS experiments is to align output data with respect to both RT and m/z.

Time and m/z axes can be aligned independently or simultaneously, though the latter has not been reported in the literature and would be more easily applied after peak detection. Furthermore, if aligning in time only, one may wish to use scalar time series rather than the vector time series most readily available from the data (e.g., total ion current (TIC) as scalar time series versus a vector of all m/z values at each time point), or even more general representation schemes, such as a reduced-dimensionality vector time series as obtainable for example by PCA.

Two main groups of existing methods for datasets alignment can be distinguished.

The first group align raw spectrum data before peak detection. These methods search for optimal warping functions to map RT of one experiment to that of another. Since the warping function only accounts for “global” variation in RT, these methods may not always align individual peptides.

The second group of alignment methods use the detected feature lists, and allow some variation in RT of individual peptides. However, since this method relies on the detected peak and does not take advantage of the raw spectrum information, the alignment decisions are vulnerable to inaccuracy in the peak detection step. In addition, both groups of methods are formulated to work on data sets that are similar to each other, and may produce bias when analyzing different samples, such as cancer and non-cancer serum. In order for LC-MS-based analysis to become a routine procedure in biomedical research, a computationally efficient and robust alignment procedure must be developed.

“Peptide Element Alignment” (PETAL) [44] uses both raw spectrum data and peak detection results to simultaneously align features from multiple LC-MS experiments. PETAL first creates spectrum elements to represent the relative intensity profiles of individual peptides. It then models the variation in retention time and the instrument noise in intensity measurements that produce error in the m/z values. Peptides detected in different LC-MS data are aligned if they are represented by the same element. By considering each peptide separately, this method offers greater flexibility than simply matching retention time between profiles. In addition, PETAL treats all experiments symmetrically and avoids the possible biases that may result from choosing one experiment as a template. Actually most algorithms used to date require a template, specified a priori, to which all time series are pre-aligned: suboptimal template choice could result in poor alignments, thus it would be wise to avoid that way.

DATA NORMALIZATION

High-throughput mass spectrometry technology offers a powerful means of analyzing biological samples. The ability of MS to identify and precisely quantify thousands of proteins from complex samples is expected to broadly affect biology and medicine. However, MS systems are subject to considerable noise and variability that is not fully characterized or accounted for. Thus, as we have just described, it is important and necessary to properly conduct data pre-processing steps such as signal filtering, peak detection, alignment in time and mass charge ratio, and amplitude normalization before reliable conclusions can be made from the data.

Since MS signals are frequently corrupted by either systematic or sporadic changes in abundance measurements, overall peak amplitudes measured in one replicate may be elevated with respect to another, and may also have systematic changes within an experiment, across time, due to a change in column or ESI performance. In such cases, the data need to be normalized to make the experiments comparable.

Furthermore, in many MS experiments, the instrument may have trouble detecting the weak signals of low-abundance peptides. Even if the instrument detects the signal, the peak intensities may be too low to be distinguished from background noise during data processing. Therefore, the lower the ion abundance, the more likely the peptide will be “missing” in the MS output data. Ignoring such non-random missing pattern may introduce significant bias into subsequent analyses: Wang et al. proposed [45] a novel probability model to describe the missing behaviour, which accounts for this type of intensity-dependent missing events.

The simplest and classical approach to normalization would be to multiply all abundance values in one experiment by some constant factor, but in general, it may be necessary to perform more detailed corrections. Normalization of MS data can be performed either by coercing m/z intensity values to be comparable across experiments (low-level processing), or by altering peak abundance to be comparable (mid-level processing). In general, one aims to normalize not only replicates, but also experimental data of distinct biological origin, such as serum profiles from cancer patients and healthy case controls. The underlying assumption behind normalization is that the overall MS abundance of either all features (peaks or time-m/z pairs), or subsets of these, should be equal across different experiments. Given this assumption, one can determine the ratio of overall abundance of a chosen set of features between two experiments for use as a multiplicative correction factor, and then normalize an entire set of experiments by arbitrarily choosing one of them as a reference to which all others are normalized, obviously such an approach is biased and error-prone.

Global normalization refers to cases where all features are simultaneously used to determine a single normalization factor between two experiments; by globally normalizing signal intensities across multiple samples, we aim to identify and remove systematic variation arising because of differential amounts of sample loaded into the LC-MS system, protein degradation over time, or variation in the sensitivity of the instrument detector. It is natural to assume that the sample intensities are all related by a constant factor. A common choice for this re-scaling coefficient is the sample mean or median. This choice is based on the assumption that the number of features whose measurements change is few compared to the total number of features. Therefore, the distribution of the measurements of all the features should be roughly the same across different experimental runs. However, in MS experiments, because of the limitation of detector sensitivity and the unavoidable instrument noise, ions below a certain intensity level may hardly be detected, which leads to non-random missing of peptide features in the result. Thus, it is not appropriate to use overall mean or median for re-scaling. In order to avoid the possible bias due to non-random missing events, it is possible to use the top N ordered statistics of feature intensities in each sample, where N is a parameter chosen by user, but this choice can be misleading [45].

Local normalization, instead, refers to cases where a subset of features are used at a time (different subsets for different parts of the data). Locality can be defined by, say, similarity in m/z values, time (scan headers), or abundance (peak intensity) levels. For example, in an abundance-dependent, local normalization, peaks of similar abundance within the same MS experiment would be scaled in a similar way, while peaks of different abundance are scaled in a different way. If the mean of all features is made to agree across all experiments, it is referred to as a global mean normalization. While several groups have opted for global abundance normalization, in the case of LC-MS data it may be necessary to normalize locally in time [46], because chromatography can produce irregular fluctuations in signal.

Many of the normalization techniques applicable to LC-MS data have also been applied to the results of microarray experiments [47]. With gene expression profiles, the genes used for normalization have sometimes been restricted to so-called “housekeeping” genes presumed to remain constant across the experimental conditions. An analogous concept was applied to LC-MS data by Wang et al. [40], whereby a constant intensity ratio between pairs of experiments was computed based on reference peaks. These authors noted, however, that the use of all detected peaks provided similar results. Moreover it is very difficult to find stable peaks (i.e., ”housekeeping”) across experiments.

Normalization is often evaluated by calculating the coefficient of variation (CV) between peaks across different experiments after normalization. While reasonable CVs (e.g., 30%) are commonly reported, a comparison to CVs from pre-normalized data is often not provided. Moreover, because no systematic comparison of these various normalization techniques has been reported, it is difficult to assess their relative merits.

4 BIOINFORMATICS CHALLENGES

In this chapter are described the main bioinformatics open issues related to the quantification and handling of quantitative mass spectrometry-based proteomics data. In particular, we focus on profile LC-MS data, which are considered the most appropriate data for quantification aims [23]. Even though quantitative MS-based proteomics significantly progressed, it is undermined by the lack of reliable bioinformatics methodologies and tools for the storage and analysis of experimental data. In fact, without efficient bioinformatics tools, high-throughput proteomics data handling and analysis could be difficult and error-prone. Expert manual analysis is incompatible with the tens of thousands of spectra collected in a single experiment and is inconsistent. Moreover, the data hostage held by different instrument proprietary formats slows down the evolution of proteomics, mainly because comparisons among different experiments, or analytical methods, often become unfeasible. These comparisons depend critically on transparent file structures for data storage, communication and visualization. Only once suited tools are tested, validated and widely accepted it will become feasible to apply quality standards for protein identification, quantification and other measurements and to compare complementary proteomic datasets generated in different laboratories.

At the state of art, most of the biological content held by the sample cannot be accessed, either for technical limitations or data analysis issues: ionization sources and mass spectrometry sensitivity needs to be enhanced; whereas the search engines (e.g., Mascot (*4*), Sequest (*5*),…) identification efficiency is commonly limited to a 30% of all peptides belonging to the sample, hence search engines and databases need to be improved (*6*); furthermore, the quantification efficiency (i.e., the number of quantified peptides) further reduces the proteome coverage at the quantification level (see Figure 4-1).

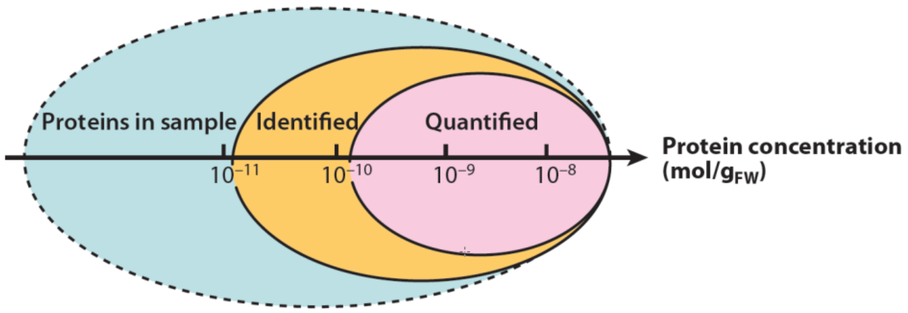


FIGURE 4-1 PROTEOME COVERAGE AT A GLANCE: USUALLY, ONLY THE HIGHER ABUNDANCE PROTEINS ARE COVERED BY IDENTIFICATION AND QUANTIFICATION.

Therefore, a major challenge facing proteomic research is how to manage the overwhelming amount of data in order to extract the qualitative and/or quantitative information on proteome and still to keep down computational costs both for data handling and processing. This holds especially for quantitative proteomics, since, in order to achieve reliable quantifications, it needs highly informative profile data, such as profile Liquid-Chromatography (LC) MS ones.

4.1 DATA HANDLING

The *Human Proteome Organization* (*HUPO*) established the *Proteomics Standards Initiative* (*PSI*) with the aim of enhancing data data comparison, exchange and verification. Established in April 2002 as a working group of the HUPO, the PSI aims to define community standards for data representation in proteomics to overcome the current fragmentation of proteomics data and to facilitate data comparison, exchange and verification. The vast amount of data associated with a single experiment can become problematic at the point of publishing and disseminating results. Only by comparing separate experiments (e.g., cells at different states, tumour cells versus normal cells) precious information concerning complex diseases can be unrevealed. Fortunately, the community has recognized and tackled the problem through the development of standards for the capturing and sharing of experimental data. The need for common formats to allow data exchange between both public and commercial database systems was recognized, as was a growing need for the establishment of public data repositories in which the ever increasing amount of published data can be deposited and retrieved by scientists working in the field and wishing to further analyze this information.

Inherent differences in the use of a variety of instruments, different experimental conditions under which analyses are performed, and potential automatic data pre-processing steps by the instrument software can influence the actual measurements and therefore the results after processing. Processing steps 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). A score, rank or confidence measure can be assigned to the result of the processing. Additionally, most instruments output has a very specific and often proprietary format. These proprietary formats are then typically transformed into so-called peak lists to be analysed by identification and characterisation software. Data reduction such as peak centroiding and deisotoping is often performed during this transformation from proprietary formats to peak lists. In addition, 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 peak lists are then used as inputs for subsequent analysis. 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.

HUPO-PSI released the *Minimum Information About a Proteomics Experiment* (*MIAPE*) reporting guidelines [48] in an effort to define the minimum set of information about a proteomics experiment that would be required by the community to share their work. The overall MIAPE standard is composed of several parts, subject to ongoing development, that describe steps for the sample processing before entering the mass spectrometer (gels, chromatography, etc.), information about the specific mass spectrometer used and the settings and results for the database searches [49]. Some of these consist only of working drafts which can be rapidly changed. As well as the MIAPE standard, large repositories for proteomics data have emerged, for example the Proteome Experimental Data Repository (PEDRo) [50], the PRoteomics IDEntifications database (PRIDE) [51,52], the Peptide Atlas database [53-55] and, lastly SRM Atlas [56]. There is indeed a need for public repositories that contain information from whole proteomics experiments; making explicit both where samples came from, and how analyses of them were performed (see Figure 4-2, borrowed from [48]). Proteomics data should therefore ideally be accompanied by contextualizing 'metadata' (essentially 'data about the data'), making explicit both where samples came from and how analyses were performed. MIAPE was preceded by the "minimum information about a microarray experiment" (MIAME) guidelines [57], which deal specifically with transcriptomics data. The microarray community similarly defined the critical information necessary to effectively describe a microarray experiment. MIAME has become an accepted community standard; the original paper had been cited in >1,100 published papers (source: Google Scholar), many of which describe MIAME-compliant software development.

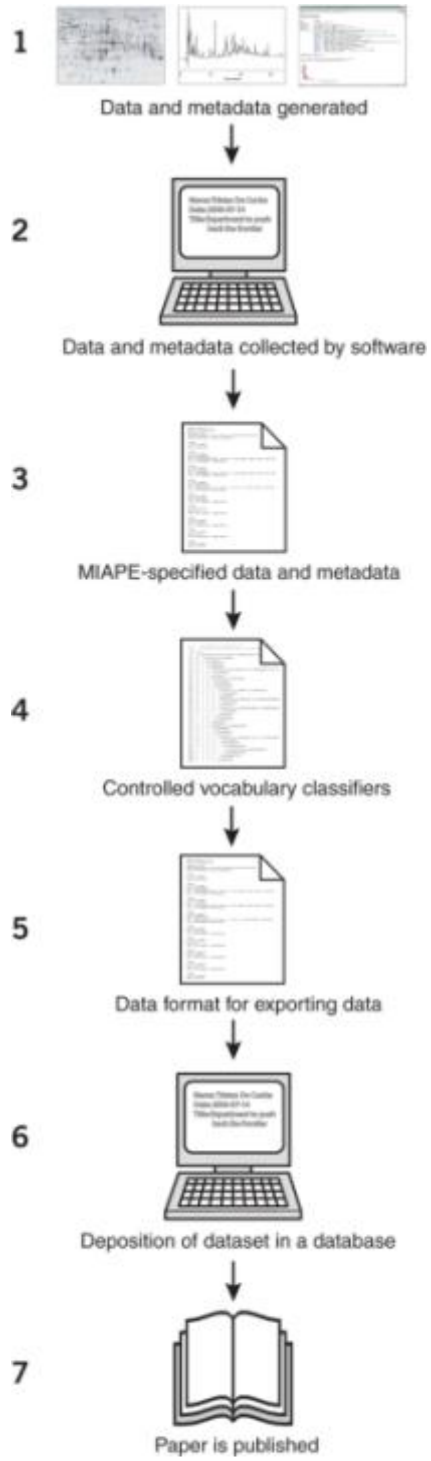

1
Data and metadata generated
2
Data and metadata collected by software
3
MIAPE-specified data and metadata
4
Controlled vocabulary classifiers
5
Data format for exporting data
6
Deposition of dataset in a database
7
Paper is published

FIGURE 4-2 (1) DATA AND METADATA ARE GENERATED BY AN EXPERIMENT; (2) SOFTWARE COLLECTS THE DATA AND METADATA, EITHER BY IMPORTING FROM COMPUTER-CONTROLLED INSTRUMENTS OR FROM MANUAL DATA ENTRY; (3) MIAPE SPECIFIES THE DATA AND METADATA TO BE REQUESTED BY THE SOFTWARE TOOL; (4) A CONTROLLED VOCABULARY SUPPLIES CLASSIFIERS VIA THE SOFTWARE; (5) THE SOFTWARE USES A DATA FORMAT SPECIFICATION WHEN EXPORTING A MIAPE-COMPLIANT DATASET; (6) THE DATASET IS STORED IN A MIAPE-COMPLIANT DATABASE AND ASSIGNED AN ACCESSION NUMBER; (7) A PAPER, INCLUDING THE APPROPRIATE ACCESSION NUMBER, IS PUBLISHED IN A JOURNAL.

4.1.1 STANDARD DATA FORMATS

The PSI-Mass Spectrometry Standards working group defines community data formats and controlled vocabulary terms facilitating data exchange and archiving in the field of proteomics mass spectrometry. They proposed *mzData* [58], which, as *mzXML* [59], is an *eXtensible Markup Language* (*XML*) [60] based data format, developed to uniform data. mzData was developed by PSI-MSS, whereas mzXML was developed at the Seattle Proteome Center (SPC) at the Institute for Systems Biology (ISB). It is recognized that the existence of two separate formats for essentially the same thing generates confusion and extra programming effort. In order to overcome the competition between them, the PSI, with full participation by ISB, recently introduced *mzML* as a unique data format [61], merging the best features of each of these formats. Finally they kept on developing a controlled vocabulary, MS CV, to be used with the previous file formats. XML-based data formats are characterized by intuitive language and a standardized structure. Here below all of them will be briefly described referring to their specifications documentation, publicly released by the HUPO-PSI (see Appendix A).

4.1.1.1 MZDATA

The mzData standard captures mass spectrometry output data as peak list information. mzData is an XML format for representing mass spectrometry data in such a way as to completely describe the instrumental aspects of the experiment (see Figure 4-3). The key feature of the format is the use of external controlled vocabularies to allow data from different instruments and experimental designs to be shared in a common format. mzData's aim was to unite the large number of current formats into a single format. It is not a substitute for the raw file formats of the instrument vendors. Some vendors, if not all, provide software transforming their raw files to mzData. There are already a number of programs which can use mzData. The format is extensible to allow the description of new instrument types; however, only mass spectrometers were included in its final and last documentation release.

mzData was meant to also be able to hold MIAPE information related to MS experiments. Parameters in mzData are stored using a generic parameter type which allows the use of either a controlled vocabulary term (cvParam) or a user defined term (userParam). The cvParam element must contain a term which is a member of a controlled vocabulary named in a cvLookup element. User-specified parameters are generic name-value elements with no reference to a formal controlled vocabulary.

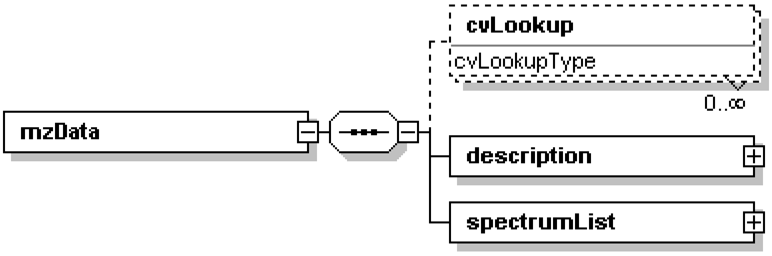


FIGURE 4-3 THIS SCHEMA CAN CAPTURE THE USE OF A MASS SPECTROMETER, THE DATA GENERATED, AND THE INITIAL PROCESSING OF THAT DATA (TO THE LEVEL OF THE PEAK LIST). PEAK LISTS ARE PROCESSED DATA FROM A MASS SPECTROMETRY EXPERIMENT.

In order to keep the file size of mzData limited, mzData format requires primary data (m/z and intensity) to be represented as base64 encoded binary using the W3C Schema base64Binary type [62]. To use this type, additional information is needed to decode the array properly (see Figure 4-4).

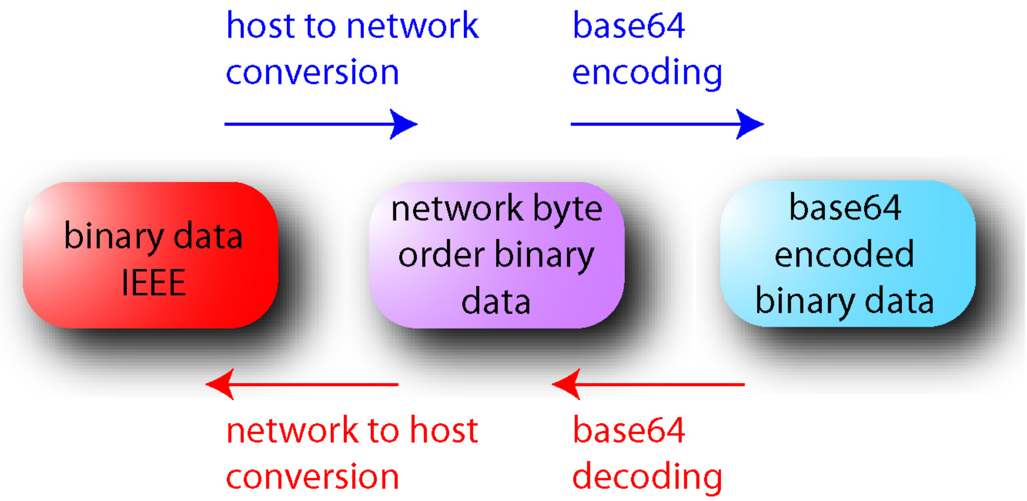

host to network conversion
base64 encoding
binary data IEEE
network byte order binary data
base64 encoded binary data
network to host conversion
base64 decoding

FIGURE 4-4 THE STEPS TO STORE DATA ARE REPRESENTED IN BLUE; THE STEPS TO EXTRACT DATA ARE REPRESENTED IN RED.

The mzData format encapsulates binary data in an element called “data”. Only IEEE-754 floats are allowed in this element, however either the 32-bit or 64-bit precision floating point representation may be used. To improve cross platform interoperability, both byte orders are allowed with order specified in the “endian” attribute. Finally, the number of floating point numbers stored in the encoded array is specified in the “length” attribute.

mzData has been released and is stable at version 1.05. It is now deprecated in favour of mzML, the current standard data format.

4.1.1.2 MZXML

mzXML is an open data format for storage and exchange of mass spectroscopy data, developed at the Seattle Proteome Center, at the Institute for Systems Biology. mzXML provides a standard container for MS and MS/MS proteomics data or multiple mass spectrometric (MSn) data, based on XML. Raw, proprietary file formats from most vendors can be converted to the open mzXML format.

XML cannot directly incorporate binary data and the conversion to a human readable clear text representation is not possible without a significant size increase. This problem is addressed in the mzXML format by encoding the (m/z, intensity) binary pairs in base64, as for mzData. As a general idea, mzXML is very similar to mzData.

The second limitation of XML representation of MS data is a consequence of some XML parsers that read a document sequentially, from the beginning of the file to the end. Therefore the mzXML schema is wrapped by a second schema, which indexes the position of each scan in a given XML file (see Figure 4-5, from [63]). At parsing time, this index can be used to adjust the input stream to a scan-specific offset.

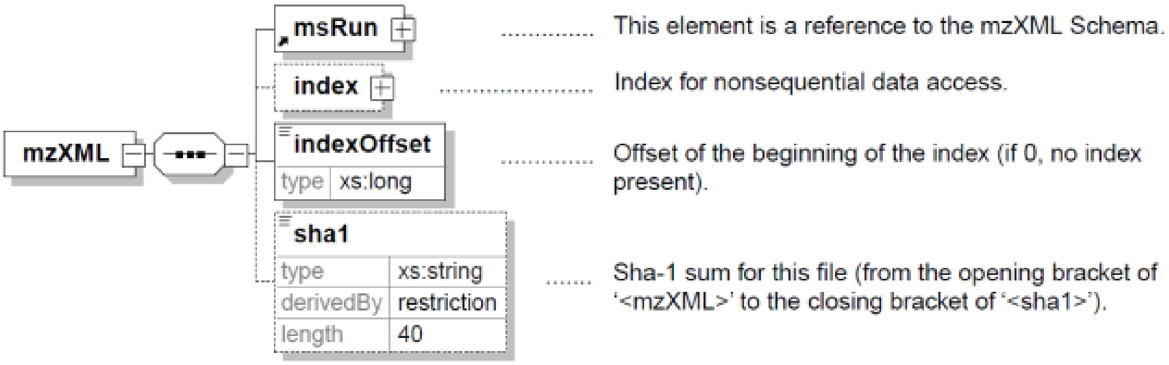

mzXML
msRun
index
indexOffset
type xs:long
sha1
type xs:string
derivedBy restriction
length 40
This element is a reference to the mzXML Schema.
Index for nonsequential data access.
Offset of the beginning of the index (if 0, no index present).
Sha-1 sum for this file (from the opening bracket of '<mzXML>' to the closing bracket of '<sha1>').

FIGURE 4-5 SCHEMA FOR THE INDEXED MZXML FORMAT. IN THE MZXML FORMAT THE ACCESS IS ADDRESSED BY INDEXING THE POSITION OF EACH SCAN IN THE DOCUMENT. THEREFORE SPECTRUM DATA CAN BE ACCESSED BY THE SCAN NUMBER. NO ACCESS INDEXING ON THE M/Z DIMENSION IS PROVIDED.

The following is a focus on the most relevant part of the mzXML schema represented in Figure 4-6. The ‘parentFile’ element stores a chronological list of all files used to generate a given instance document. The ‘msInstrument’ element stores the specifications of the MS instrument (e.g., resolution, manufacturer, model, ionization type, mass analyzer type, detector type) and acquisition software used to generate the data. A ‘nameValue’ element provides a means to store laboratory-specific instrument modifications. Even in a vendor-neutral representation, it is important to preserve this information because the analytical software should account for the strengths and weaknesses of different instruments. The ‘dataProcessing’ element describes any type of data processing (e.g., centroiding, noise reduction, peak finding) performed during the creation of the current instance document. The ‘scan’ element has attributes to describe, among others, the retention time, the MS level, the polarity of the ion source, the ionization energy and the mode of acquisition (e.g., full, selected ion monitoring, selected reaction monitoring) for the scan being described. The ‘scan’ element contains a reference to itself, which provides an intuitive way to store scans sharing a common ancestor (e.g., a common survey scan). It features seven sub-elements: the ‘scanOrigin’ element, the ‘precursorMz’ element, the ‘maldi’ element, the ‘peaks’ element, the ‘nameValue’ element and the ‘comment’ element (for an example see Figure 4-7).

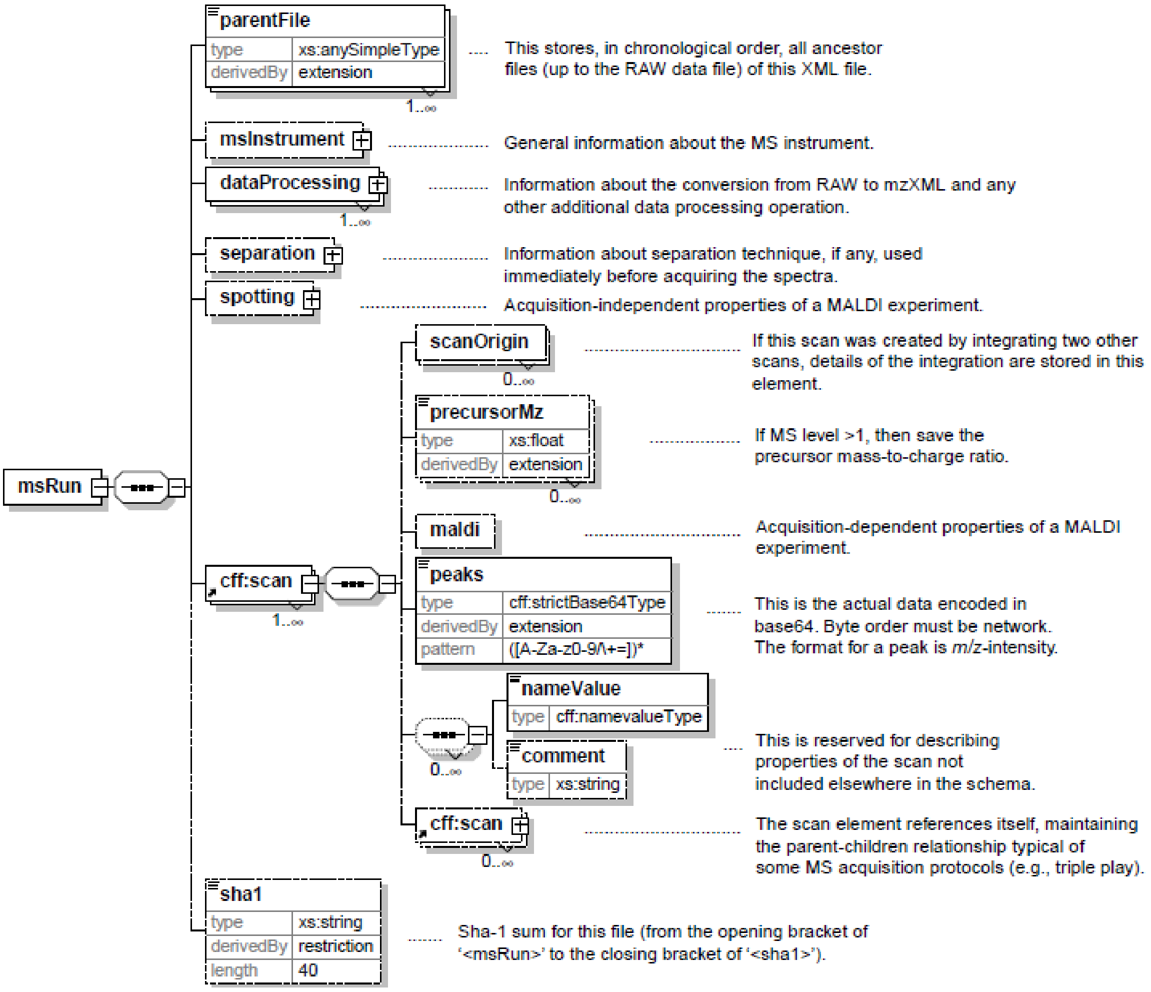
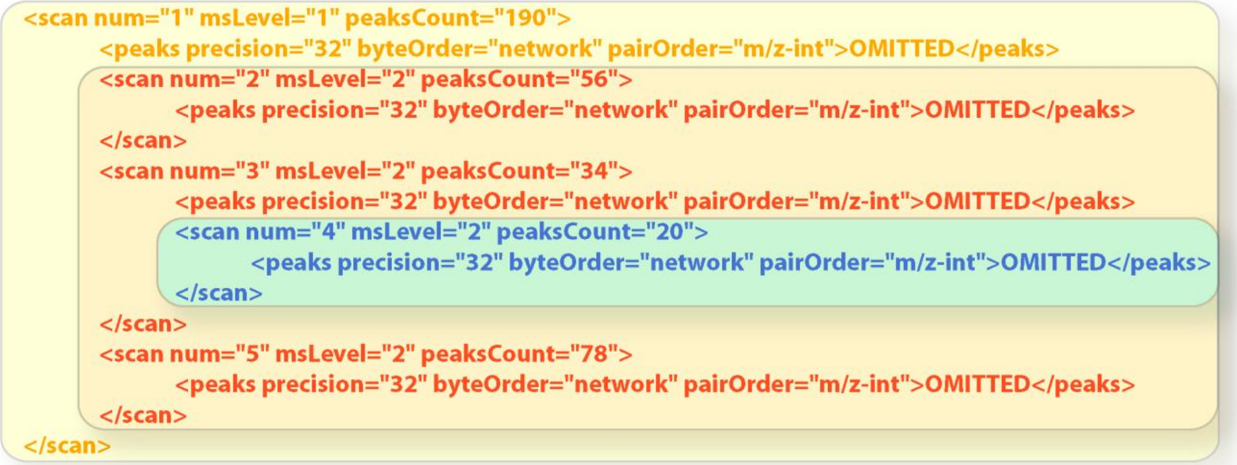

parentFile
type xs:anySimpleType
derivedBy extension
1..∞
This stores, in chronological order, all ancestor files (up to the RAW data file) of this XML file.
msInstrument
General information about the MS instrument.
dataProcessing
1..∞
Information about the conversion from RAW to mzXML and any other additional data processing operation.
separation
Information about separation technique, if any, used immediately before acquiring the spectra.
spotting
Acquisition-independent properties of a MALDI experiment.
msRun
scanOrigin
0..∞
If this scan was created by integrating two other scans, details of the integration are stored in this element.
precursorMz
type xs:float
derivedBy extension
0..∞
If MS level >1, then save the precursor mass-to-charge ratio.
maldi
Acquisition-dependent properties of a MALDI experiment.
cff:scan
1..∞
peaks
type cff:strictBase64Type
derivedBy extension
pattern ([A-Za-z0-9/\+=])*
This is the actual data encoded in base64. Byte order must be network. The format for a peak is m/z-intensity.
nameValue
type cff:namevalueType
comment
type xs:string
0..∞
This is reserved for describing properties of the scan not included elsewhere in the schema.
cff:scan
0..∞
The scan element references itself, maintaining the parent-children relationship typical of some MS acquisition protocols (e.g., triple play).
sha1
type xs:string
derivedBy restriction
length 40
Sha-1 sum for this file (from the opening bracket of '<msRun>' to the closing bracket of '<sha1>').

FIGURE 4-6 OVERVIEW OF THE MZXML SCHEMA VERSION 2.0. THIS VERSION IS COMPATIBLE WITH LC-ESI-MSN AND WITH MALDI-MSN EXPERIMENTS.

The 'scanOrigin' sub-element stores the details of the integration process if the current scan has been created by merging multiple scans. The 'precursorMz' sub-element stores the m/z, intensity, charge state, width of the selection window and collision energy values for the precursor ion fragmented in the current scan. Multiple instances of the 'precursorMz' sub-element per scan element can be included to account for fragmentation spectra possessing more than one precursor ion (e.g., as in shotgun sequencing experiments with fragments generated by in-source decay [64]). The 'maldi' sub-element stores those parts of data from a MALDI experiment that can vary between multiple scans acquired on the same spot (e.g., the laser intensity or the duration of the laser excitation). The 'peaks' sub-element contains the m/z intensity pairs as base64-encoded binary data. This element can store raw as well as processed m/z intensity pairs. The ‘nameValue’ sub-element can be used to add entries to the instance document without having to change the schema. This allows different laboratories to have personalized instance documents, while referring to a centralized common schema.



```
<scan num="1" msLevel="1" peaksCount="190">
        <peaks precision="32" byteOrder="network" pairOrder="m/z-int">OMITTED</peaks>
        <scan num="2" msLevel="2" peaksCount="56">
                <peaks precision="32" byteOrder="network" pairOrder="m/z-int">OMITTED</peaks>
        </scan>
        <scan num="3" msLevel="2" peaksCount="34">
                <peaks precision="32" byteOrder="network" pairOrder="m/z-int">OMITTED</peaks>
                <scan num="4" msLevel="2" peaksCount="20">
                        <peaks precision="32" byteOrder="network" pairOrder="m/z-int">OMITTED</peaks>
                </scan>
        </scan>
        <scan num="5" msLevel="2" peaksCount="78">
                <peaks precision="32" byteOrder="network" pairOrder="m/z-int">OMITTED</peaks>
        </scan>
</scan>
```

FIGURE 4-7 MZXML INSTANCE DOCUMENT IF THE MS INSTRUMENT WAS SET TO DO ONE MS SURVEY SCAN (YELLOW) FOLLOWED BY 3 MS/MS SCANS (RED) AND ONE MS/MS/MS SCAN (LIGHT BLUE) SELECTED FROM THE SECOND MS/MS SCAN.

4.1.1.3 MZML

mzML is a new format which aims to merge the best elements of mzXML and mzData, and represents a joint effort of the HUPO/PSI committee, SPC/ISB, instrument vendors, and other proteomics software groups. mzML is intended to replace all earlier formats. mzML is a common open format to record the output of mass spectrometers prior to database searching or other downstream processing of the spectra. It is designed to hold the data output of a mass spectrometer as well as a systematic description of the conditions under which this data was acquired and transformed. The mzML schema is designed to contain all the information for a single MS run, including meta data about the spectra plus all the spectra themselves, either in centroided (peak list) or profile mode. The primary focus of the model is to support long-term archiving and sharing, rather than day-to-day laboratory management, although the model is extensible to support context-specific details. In order to properly describe mass spectrometry data output and the experimental context mzML includes: the actual data acquired, to a sufficient precision, as well as its associated metadata; and an adequate description of the instrument characteristics, its configuration and possible pre-processing steps applied.

The header at the top of the file encodes information about the source of the data as well as information about the sample, instrument and software that processed the data. The element <mzML> is the root element for the Proteomics Standards Initiative (PSI) mzML schema, which is intended to capture the use of a mass spectrometer, the data generated, and the initial processing of that data (to the level of the peak list). The element <spectrum> captures the generation of a peak list (including the underlying acquisitions). Also describes some of the parameters for the mass spectrometer for a given acquisition (or list of acquisitions). The mzML specification also supports the <chromatogram> element, which is very similar to the <spectrum> element. It is capable of containing a full description of and the data for a chromatogram. The chromatogram may be simply be a total ion current (TIC) chromatogram of an ordinary MS1 or MS/MS run, or a chromatogram corresponding to a Q1,Q3 pair in a SRM run. Selected reaction monitoring (SRM) is the major new technology that is supported by mzML that was not supported by both previous formats. There was considerable discussion on how to encode SRM experiments: as tiny MS/MS-like spectra or directly as complete chromatograms. The decision was made that each SRM scan is to be encoded as a mini MS/MS-like spectra with a precursor corresponding to the Q1 m/z and a small spectrum encoding one or more Q3 m/z values that correspond to the Q1 m/z. We note that these mini scans may be a single (centroided) value per Q3 m/z, or the mini scans may be profile mode scans surrounding the Q3 m/z. For example, it is entirely permissible to monitor two Q3 m/z values for a single Q1 m/z, and encode profile mode scans for both Q3 regions in a single spectrum. It has been resolved that all SRM runs must be encoded as mini MS/MS-like spectra using the <spectrum> element. Optionally, the same information may also be encoded using the <chromatogram> elements as a speed-enhancing feature. At present, it has been decided that SRM output may not be encoded only in the <chromatogram> form. The goal is to avoid having two different ways of encoding the same data. Readers can always count on the mini MS/MS-like spectra and may only optionally support the <chromatogram> constructs. This was merely a policy decision, not one dictated by the schema. The mzML model is described in the XML schema showed in Figure 4-8.

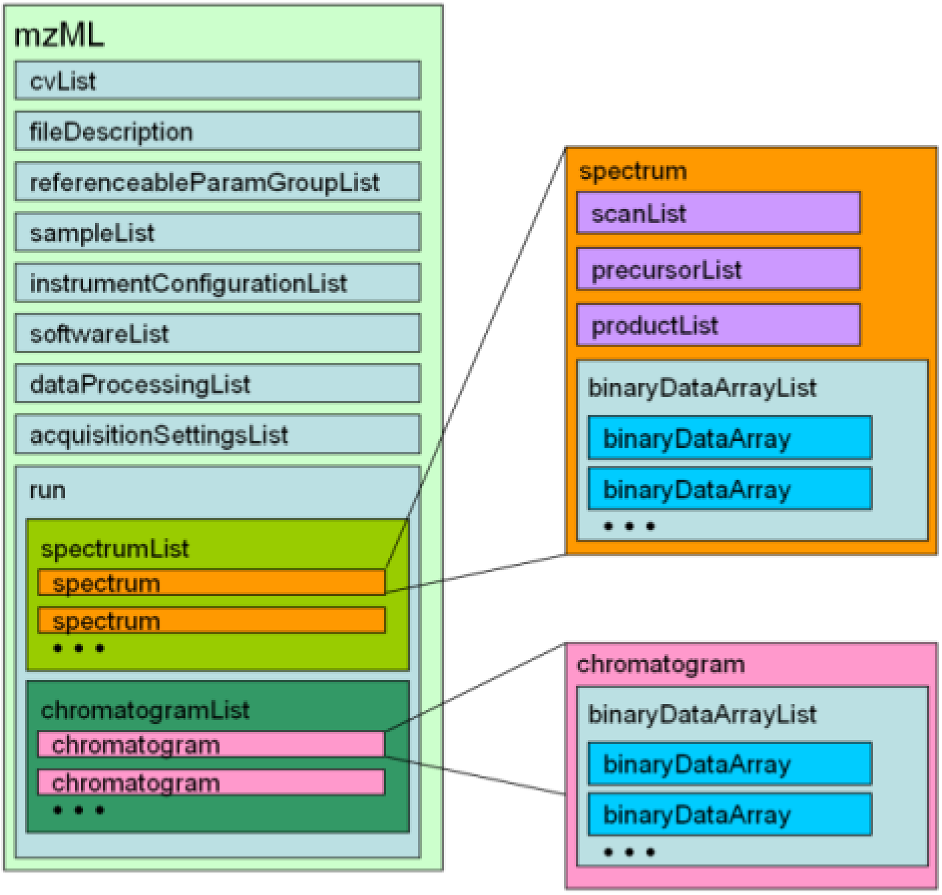


FIGURE 4-8 HIGH LEVEL OVERVIEW OF THE XML ELEMENTS FOR MZML. EACH BOX REPRESENTS AN XML ELEMENT, NESTED WITHIN OTHER ELEMENTS AS SHOWN. MZML MAY BE ENCLOSED IN A SPECIAL INDEXING WRAPPER SCHEMA TO ALLOW RANDOM ACCESS INTO THE FILE, ALLOWING SOFTWARE TO PULL OUT ONE OR MORE ARBITRARY SPECTRA. EACH SPECTRUM CONTAINS A HEADER WITH SCAN INFORMATION AND OPTIONALLY PRECURSOR INFORMATION, FOLLOWED BY TWO OR MORE BASE64-ENCODED BINARY DATA ARRAYS. CHROMATOGRAMS MAY BE ENCODED IN MZML IN A SPECIAL ELEMENT THAT CONTAINS ONE OR MORE CVPARAMS TO DESCRIBE THE TYPE OF CHROMATOGRAM, FOLLOWED BY TWO BASE64-ENCODED BINARY DATA ARRAYS.

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 xsd 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.

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.

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 will resolve the current problems.

One of the aspects of mzXML that enabled its very swift adoption was a ready set of open source tools that implemented the format. With these tools many users were able to immediately begin using the format without coding their own software. Therefore, to insure that mzML is a format that will quickly be adopted and implemented uniformly, the format is presented with several tools that write, read, and validate the format.

The byte-offset index that allowed random access to arbitrary spectra within the file was retained for mzML. mzML documents themselves do not have an index. A reference implementation is provided for indexing as a wrapper schema for an mzML document.

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. This is evidenced by mzData still being at version 1.05 while mzXML is currently at version 3.1. 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 should be solved for mzML by releasing a semantic validator with the data format (see Figure 4-9). 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 of using the semantic validator, not just an XML validator. 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 a curse.

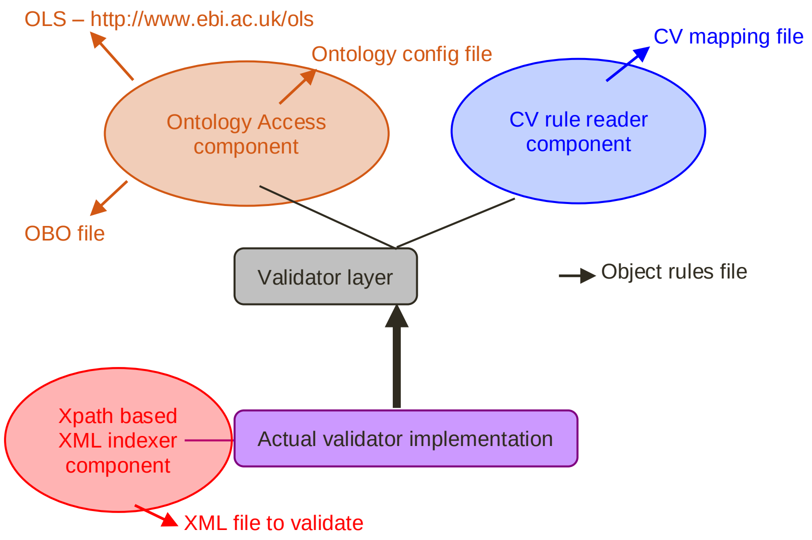
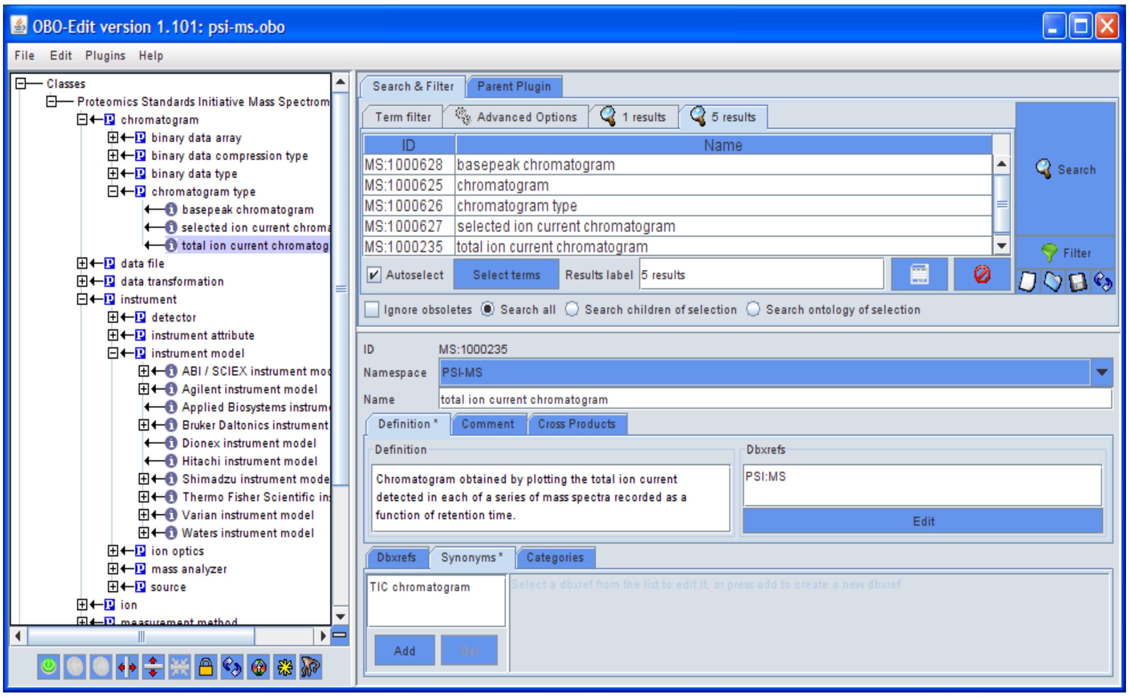

OLS – http://www.ebi.ac.uk/ols
Ontology config file
Ontology Access component
OBO file
CV mapping file
CV rule reader component
Validator layer
Object rules file
Xpath based XML indexer component
Actual validator implementation
XML file to validate

FIGURE 4-9 A SCHEMATIC REPRESENTATION OF THE SEMANTIC VALIDATOR. IT IS AVAILABLE AS A WEB PAGE (HTTP://EDDIE.THEP.LU.SE/PRODAC_VALIDATOR/VALIDATOR.PL) OR AS A STANDALONE TOOL.

A comprehensive collection of terms have been defined (mostly extracted from vocabulary and definitions of the IUPAC nomenclature book [65]) and structured in an mzML-friendly way, hopefully facilitating the browsing of the terms. Almost all first-level branch terms (the direct children of the root term) have a homonymous XML element in mzML. Their children, the second-level terms, are relevant topics or categories which need CV support for their description. The leaf nodes under their respective parent categories should be used in a cvParam (further details in Figure 4-10) under the appropriate XML element in mzML schema.

Although the structure of the CV and the mzML schema are related, the details of which terms are allowed/recommended in a given schema section is reported in the mapping file. The mapping file is a list of associations between a cvParam element in a specific schema and the branches of the CV terms expected in that location. This file is read and interpreted by the validator, checking that the data annotation is consistent. The mapping file needs to be checked and eventually updated when the CV terms or structure are changed.



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 * | Comment | Cross Products

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

Dbxrefs | Synonyms * | Categories

TIC chromatogram

FIGURE 4-10 MUCH OF THE METADATA ENCODED IN THE MZML IS IN THE FORM OF CVPARAMS, 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 AND IS READ IN BY MOST READERS AND WRITERS OF MZML. THE CONTROLLED VOCABULARY CAN BE EASILY ADJUSTED AND EXTENDED WITHOUT MODIFYING THE SCHEMA.

It was decided that all list elements would have a count attribute. The reason is that parsers implemented in languages where memory allocation or array sizing is important, it is a nice performance enhancement to have a count attribute indicating how many elements there are in the list. Although it was felt that this is an easy target for creating inconsistent files (i.e., writing out a count=”5” attribute followed by 6 items in the list), this was deemed to be rare and in the vast majority of cases the value can be relied on. The code would need to handle cases where the count was incorrect, but this is no more difficult than not knowing the value ahead of time.

As it has been seen before, mass spectra can be profile and centroided. Profile spectra represent the scanned data in an approximately regularly spaced format, sometimes with gaps. Centroided spectra present the scanned data only by specifying the location and intensity of individual detected peaks, usually after subjecting the profile spectrum to a peak-picking algorithm. The mzML format can encode either format with the specification of the proper controlled vocabulary term indicating which one. However, it is not allowed to encode the same spectrum in both profile and centroided modes in the same file. This is because the id attribute should nominally be the same and may not be duplicated. The recommended workflow if both spectra are desired is to encode the profile spectra in one file and the peak-picked data in a second file (with appropriate annotations as to what was done). It is permissible to have some spectra in one mode and different ones in another; for example MS level 1 spectra may be profile mode, while MS level 2 spectra may be peak picked in the same file.

4.1.2 COMPUTATIONAL ISSUES

At the state of art, the adoption of these formats is widespread among the proteomics research groups, also thanks to the extensive support of instrument and database searching vendors, and the availability of converters from proprietary data formats. In spite of their success, the currently adopted formats suffer from some limitations [63]: the impossibility to store raw data [23]; the lack of information on the experimental design, necessary for regulatory submission; the lack of scalability with respect to data size, a bottleneck for the analysis of profile data. Above all, the 1-dimensional (1D) data indexing provided by these formats considerably penalizes the analysis of datasets embodying an inherent 2-dimensional (2D) indexing structure, such as *Liquid Chromatography-MS* (*LC-MS*) ones.

Minimizing the computational time to access these huge datasets plays a key role in the progress of LC-MS data mining, and can be of help also in a variety of other MS techniques, since MS experiments usually have a “temporal” index related to the experimental time at which the MS acquisition takes place (e.g., a *scan* in mzXML). Depending on the downstream analysis, MS data can be retrieved as 2D or 3D signal by means of different accesses, based on either a m/z range, or a temporal range, or a combination of them, defining different range queries. On LC-MS data, these accesses provide respectively *chromatograms* (2D), *spectra* (2D), and *peptide* data (3D), whereas on generic MS data, they provide a set of sub-spectra belonging to the specified range. An elevated number of range queries are required during data analysis, thus an optimized data access strategy would significantly improve computational performance.

Most research groups develop, often in a sub-optimal way, intermediate data structures optimized for accesses on a privileged dimension, depending on the downstream analysis. The lack of a standard procedure for data analysis delayed the development of a standard data format optimized for computation. For instance, accredited software packages like *Maspectras* [66,67] and *MapQuant* [42] make use of the method-specific intermediate data structures *Chrom* and *OpenRaw*, respectively: the former is optimized for a chromatogram based access, the latter for a spectra based access. Chrom is a textual data format where each row stores one chromatogram. Raw data are stored in binary data files organized in three functionally distinct folders, contained within a parent folder named after the LC/MS experiment. These folders are: a global parameters folder (labeled PARAM), an MS spectra archive folder (labeled MS1), and an MS/MS spectra archive folder (labeled MS2). In a recent work [68] Khan et al. provide evidence that the use of a spatial indexing structure, namely the kd-tree, is suitable for handling large LC-MS datasets and supporting the extraction of quantitative measurements. The authors emphasize the effectiveness of the kd-tree for performing analyses based on range queries but they do not compare explicitly the range query performance of the kd-tree with that attainable by other known data structures. Moreover, their experimental assessment is carried out only on centroid datasets and does not consider profile data, which, as the literature often remarks, are the most informative [23], especially for quantitative analysis, but also the most challenging to handle. For this reason one of the objectives of this thesis was to develop a data structure to efficiently access profile data.

4.2 QUANTIFICATION

Quantification is one of the most important open issues in mass spectrometry-based proteomics [37,69-71]. Although reliable protocols are typically available to carry out the quantification from the initial samples up to the measurements on the mass spectrometer, the limiting factor in an analysis pipeline today is often found at the stage of data processing. Indeed, people often rely on software applications they do not fully understand or that provide precious little documentation or background information (the notorious black-box problem that pervades several aspects of data processing in high-throughput fields such as proteomics). As a result, users often fail to perceive correctly the strengths and limitations of their data processing tools, and the areas of application where they perform optimally. In the following, some of the most important among available software will be illustrated.

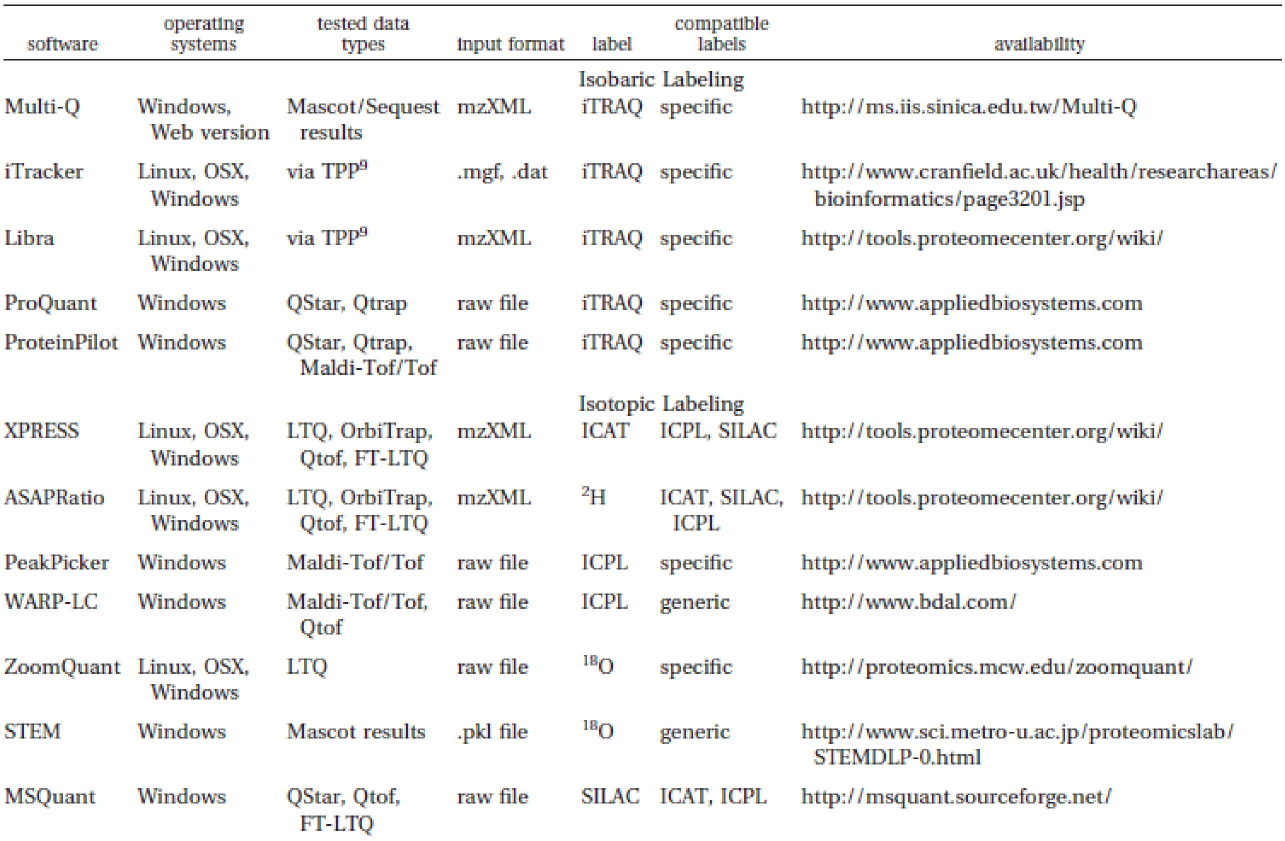
4.2.1 AVAILABLE SOFTWARE

A sizable number of software tools is now available that support quantification of LC-MS experiments. As is common to many research fields, software development is a dynamic process and proceeds in conjunction with technical advances of analytical instruments. LC-MS software tools are developed for specific generations or types of mass spectrometers and may produce high-quality results only with data generated by a limited number of MS platforms. These utilized platforms consequently define the theoretical limits of the computational LC-MS analysis (sensitivity and specificity). Therefore, it is often not trivial to choose an appropriate program suitable for the quantification of data generated by a specific instrument. Moreover, no comparison among them was ever provided by literature.

During the last decade, many research groups developed quantification software to analyze their own data. Most of these tools accept few data formats often generated by a single instrument, while data need to be produced under a strictly defined experimental workflow. Conversely, some tools have been developed for a widespread use, such as the freely available ASAPRatio [72] (embedded in the Trans Proteomic Pipeline [73]) and MaxQuant [74], or the licensed Mascot Distiller [75]. They showed good quantification performance and are commonly used among proteomic research laboratories.

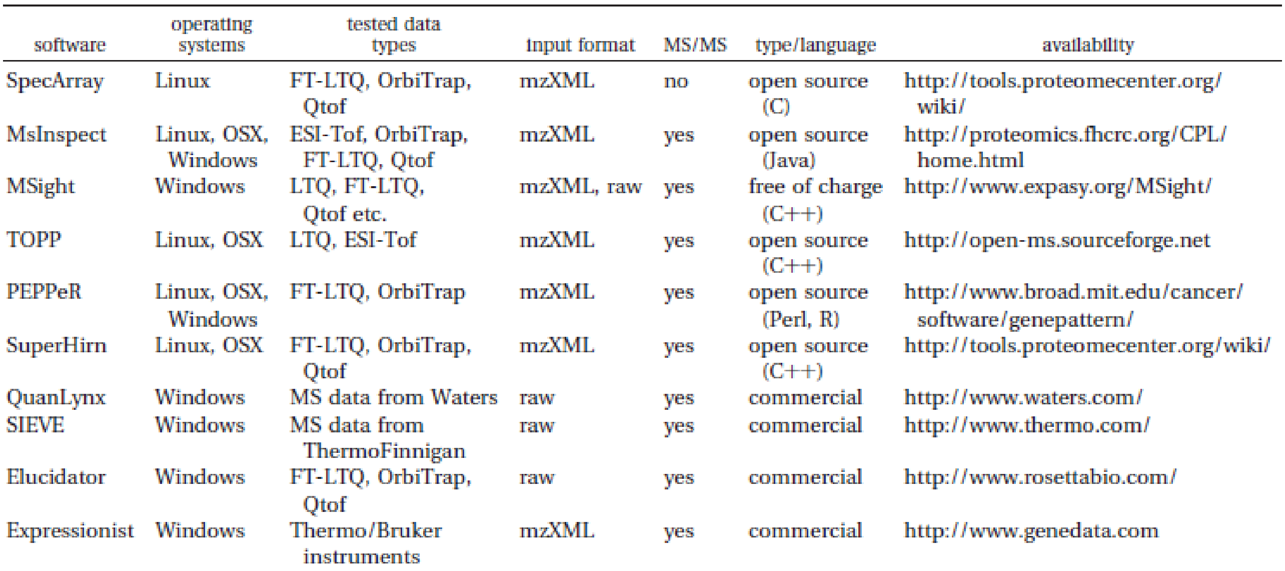
The different tools for quantitative proteomics have different strengths and weaknesses. Recently, a software called Rover [37,76], has been released, which enables to compare different quantification methods.

Figure 4-11 and Figure 4-12 are tables borrowed from [37], which report the existing software for differential and label free quantification strategies, respectively.



software	operating systems	tested data types	input format	label	compatible labels	availability
				Isobaric Labeling		
Multi-Q	Windows, Web version	Mascot/Sequest results	mzXML	iTRAQ	specific	http://ms.iis.sinica.edu.tw/Multi-Q
iTracker	Linux, OSX, Windows	via TPP[9]	.mgf, .dat	iTRAQ	specific	http://www.cranfield.ac.uk/health/researchareas/bioinformatics/page3201.jsp
Libra	Linux, OSX, Windows	via TPP[9]	mzXML	iTRAQ	specific	http://tools.proteomecenter.org/wiki/
ProQuant	Windows	QStar, Qtrap	raw file	iTRAQ	specific	http://www.appliedbiosystems.com
ProteinPilot	Windows	QStar, Qtrap, Maldi-Tof/Tof	raw file	iTRAQ	specific	http://www.appliedbiosystems.com
				Isotopic Labeling		
XPRESS	Linux, OSX, Windows	LTQ, OrbiTrap, Qtof, FT-LTQ	mzXML	ICAT	ICPL, SILAC	http://tools.proteomecenter.org/wiki/
ASAPRatio	Linux, OSX, Windows	LTQ, OrbiTrap, Qtof, FT-LTQ	mzXML	^{2}H	ICAT, SILAC, ICPL	http://tools.proteomecenter.org/wiki/
PeakPicker	Windows	Maldi-Tof/Tof	raw file	ICPL	specific	http://www.appliedbiosystems.com
WARP-LC	Windows	Maldi-Tof/Tof, Qtof	raw file	ICPL	generic	http://www.bdal.com/
ZoomQuant	Linux, OSX, Windows	LTQ	raw file	^{18}O	specific	http://proteomics.mcw.edu/zoomquant/
STEM	Windows	Mascot results	.pkl file	^{18}O	generic	http://www.sci.metro-u.ac.jp/proteomicslab/STEMDLP-0.html
MSQuant	Windows	QStar, Qtof, FT-LTQ	raw file	SILAC	ICAT, ICPL	http://msquant.sourceforge.net/

FIGURE 4-11 THE TABLE SUMMARIZES SOFTWARE PROGRAMS FOR THE QUANTIFICATION OF DIFFERENTIAL LABELING EXPERIMENTS. SOFTWARE COMPATIBILITY TO OTHER LABELING TECHNIQUES IS SHOWN WHERE A PROGRAM IS EITHER LIMITED TO A CERTAIN LABEL (SPECIFIC) OR APPLICABLE TO DIFFERENT LABELING STRATEGIES (GENERIC). FOR SOME TOOLS, THE COLUMN “COMPATIBLE LABELS” SUMMARIZES FOR WHICH ISOTOPIC LABELS THE PROGRAM HAS ALREADY BEEN TESTED.



software	operating systems	tested data types	input format	MS/MS	type/language	availability
SpecArray	Linux	FT-LTQ, OrbiTrap, Qtof	mzXML	no	open source (C)	http://tools.proteomecenter.org/wiki/
MsInspect	Linux, OSX, Windows	ESI-Tof, OrbiTrap, FT-LTQ, Qtof	mzXML	yes	open source (Java)	http://proteomics.fhcrc.org/CPL/home.html
MSight	Windows	LTQ, FT-LTQ, Qtof etc.	mzXML, raw	yes	free of charge (C++)	http://www.expasy.org/MSight/
TOPP	Linux, OSX	LTQ, ESI-Tof	mzXML	yes	open source (C++)	http://open-ms.sourceforge.net
PEPPeR	Linux, OSX, Windows	FT-LTQ, OrbiTrap	mzXML	yes	open source (Perl, R)	http://www.broad.mit.edu/cancer/software/genepattern/
SuperHirn	Linux, OSX	FT-LTQ, OrbiTrap, Qtof	mzXML	yes	open source (C++)	http://tools.proteomecenter.org/wiki/
QuanLynx	Windows	MS data from Waters	raw	yes	commercial	http://www.waters.com/
SIEVE	Windows	MS data from ThermoFinnigan	raw	yes	commercial	http://www.thermo.com/
Elucidator	Windows	FT-LTQ, OrbiTrap, Qtof	raw	yes	commercial	http://www.rosettabio.com/
Expressionist	Windows	Thermo/Bruker instruments	mzXML	yes	commercial	http://www.genedata.com

FIGURE 4-12 OVERVIEW OF LC-MS QUANTIFICATION PROGRAMS FOR LABEL-FREE QUANTIFICATION. SOFTWARE FEATURES SUCH AS PROGRAM PORTABILITY AND AVAILABILITY, DATA COMPATIBILITY, AND INTEGRATION OF MS/MS INFORMATION (MS/MS) ARE SUMMARIZED. MS/MS: IF THE SOFTWARE PROVIDES FUNCTIONALITY FOR THE INTEGRATION OF MS/MS INFORMATION. RAW: SOFTWARE IMPORTS LC-MS DATA FROM INSTRUMENT RAW FILES.

4.2.1.1 ASAPRATIO

ASAPRatio (Automated Statistical Analysis of Protein Abundance Ratios) [77] performs quantification after peptide sequence identification and verification. It collects this information from output files of the INTERACT [73] data organizing tool: peptide sequences, scan numbers, charge states at their identification, corresponding proteins and experiment data files. It has the flexibility required for the analysis of data generated from peptides labeled with multiple and diverse isotopic tags and its quantification performance has been assessed in the published results. Both peptide identifications and quantifications were manually validated. It made use of the signals recorded for the different isotopic forms of peptides of identical sequence. It performs numerical and statistical methods, such as Savitzky-Golay smoothing filters, statistics for weighted samples, and Dixon's test for outliers.

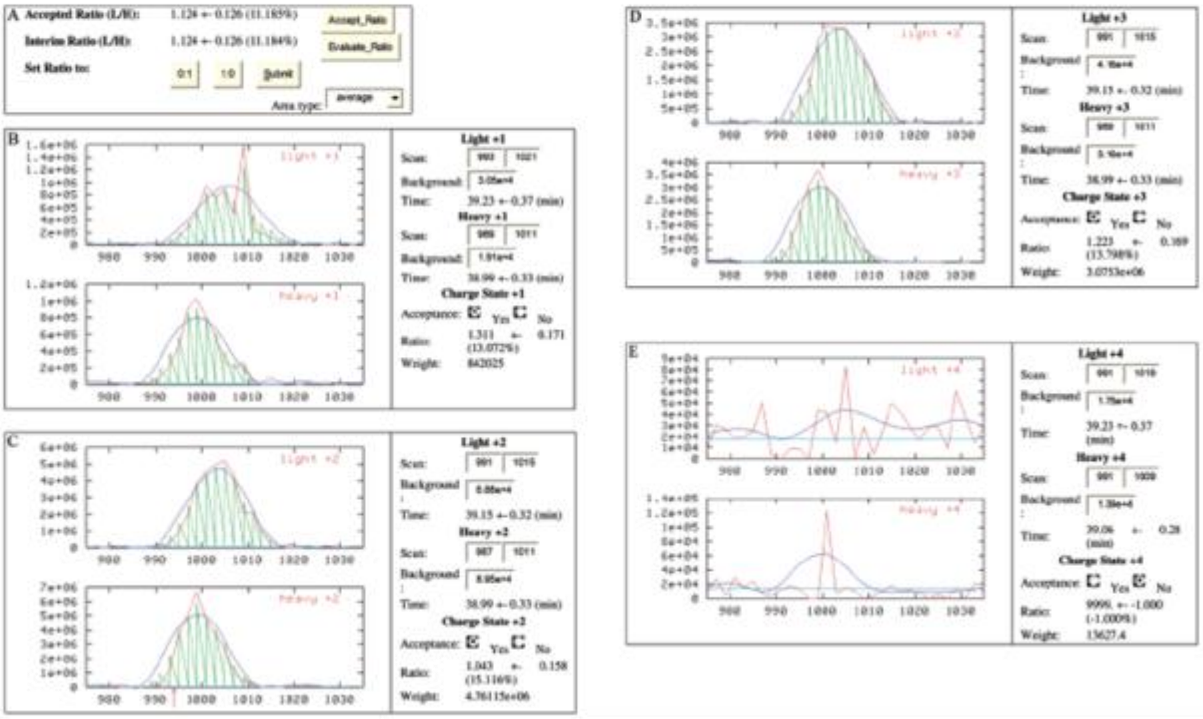
Here, ASAPRatio will be thoroughly described because it is the most important quantification method for the quantification of the data we are analyzing. Indeed, in the following, a quantification software for ICPL data will be proposed and its quantification performance will be assessed by the comparison to the one reached by ASAPRatio.

Its procedure to determine protein quantification and profiling consists of 4 steps.

Step 1 is the evaluation of a peptide abundance ratio for each peptide identified by MS/MS and database searching (see Figure 4-13).

- If both the peptide and its isotopic partner have acceptable elution peaks, an abundance ratio is calculated as the ratio of the two corresponding elution peak areas, which are calculated from the averages of the raw and the smoothed chromatograms.
- If one or both of the peak areas were set to zero, the abundance ratio is set to 1:0 or 0:1 or denoted “unquantifiable”.
- The ratio error is propagated from the area errors, which is calculated from the signal difference of the raw and the smoothed chromatograms.
- For each observed charge state, the ASAPRatio program calculates an abundance ratio. Every ratio weighted by the sum of the two corresponding elution peak areas is then used to calculate a peptide abundance ratio and its standard deviation by statistical methods for weighted samples.
- If there are at least three abundance ratios, Dixon’s test is applied to eliminate any outliers prior to statistical analysis.

The result of step one of the process is a weighted abundance ratio for each observation of an identified peptide.



A Accepted Ratio (L/H): 1.124 +- 0.126 (11.185%)

Interim Ratio (L/H): 1.124 +- 0.126 (11.184%)

Set Ratio to:

Area type: average

Light +1
Scan: 993 | 1021
Background: 3.05e+4
Time: 39.23 +- 0.37 (min)
Heavy +1
Scan: 989 | 1011
Background: 1.91e+4
Time: 38.99 +- 0.33 (min)
Charge State +1
Acceptance: ☒ Yes ☐ No
Ratio: 1.311 +- 0.171 (13.072%)
Weight: 842025

Light +2
Scan: 991 | 1016
Background: 6.88e+4
Time: 39.15 +- 0.32 (min)
Heavy +2
Scan: 987 | 1011
Background: 8.05e+4
Time: 38.99 +- 0.33 (min)
Charge State +2
Acceptance: ☒ Yes ☐ No
Ratio: 1.043 +- 0.158 (15.116%)
Weight: 4.76115e+06

Light +3
Scan: 991 | 1015
Background: 4.16e+4
Time: 39.15 +- 0.32 (min)
Heavy +3
Scan: 989 | 1011
Background: 3.16e+4
Time: 38.99 +- 0.33 (min)
Charge State +3
Acceptance: ☒ Yes ☐ No
Ratio: 1.223 +- 0.169 (13.798%)
Weight: 3.0753e+06

Light +4
Scan: 991 | 1016
Background: 1.75e+4
Time: 39.23 +- 0.37 (min)
Heavy +4
Scan: 991 | 1008
Background: 1.38e+4
Time: 39.06 +- 0.28 (min)
Charge State +4
Acceptance: ☐ Yes ☒ No
Ratio: 9999. +- -1.000 (-1.000%)
Weight: 13627.4

Red: raw signal

Blue: smoothed signal

$$R = \sum_{i=1}^{4} w_i \bullet R_i$$

$$w_i = A_i^{light} + A_i^{heavy} \qquad R_i = A_i^{light} / A_i^{heavy}$$

FIGURE 4-13 THE FIGURE ILLUSTRATES THE EVALUATION OF A PEPTIDE ABUNDANCE RATIO. R SYMBOL IS THE RATIO COMPUTED AS THE WEIGHTED SUM OF ALL ABUNDANCE RATIOS OF PEPTIDE OCCURRENCES IN DIFFERENT CHARGE STATES. THE WEIGHTS W ARE GIVEN BY THE SUMMATION OF THE PEPTIDE AREA IN THE LIGHT AND HEAVY FORM. THE RATIO FOR EACH CHARGE STATE IS THE RATIO OF THE AREA RELATED TO THE LIGHT FORM OF THE PEPTIDE TO THE AREA OF THE AREA RELATED TO THE HEAVY FORM OF THE PEPTIDE.

Step 2 is the evaluation of a "unique peptide ratio" for each identified peptide sequence.

Since in a dataset there are multiple independent observations of the same peptide, ASAPRatio evaluates peptide's contribution to the "unique peptide ratio" obtained from all the measured peptide abundance ratios of the peptide itself. Such an evaluation takes place in 2 substeps:

- Peptide abundance ratios (as in Step1) of all peptides identified during the same RP elution peak (different isotopic forms/charge states) are first grouped together to calculate an abundance ratio for the RP peak.

- Abundance ratios of different RP peaks (either in different chromatographic fractions or at different elution times during the same RP run) weighted by the areas of the corresponding RP elution peaks are used to calculate the unique peptide ratio.

If there are at least three individual ratios, Dixon’s test is applied to identify outliers. The result of this step of the process is a weighted unique abundance ratio for each identified peptide.

Step 3 is the evaluation of protein abundance ratio for each identified protein.

- Statistical methods for weighted samples are applied to calculate the protein abundance ratio and its associated standard deviation from all of its corresponding unique peptide ratios.
- The unique peptide ratios are weighted by their errors.
- If three or more unique peptides are identified for a protein, Dixon’s test is applied to identify any outlier peptides. An interface using CGI programming is available for users to verify protein abundance ratios.

The result of this step of the process is a weighted protein abundance ratio for each identified protein for which at least one peptide has been identified and quantified.

Step 4 is the evaluation of the significance of abundance change for each identified protein.

In quantitative proteomics, protein abundance ratios are typically used to identify differentially expressed proteins without considering the effect of the confidence level. This could be misleading for the identification of changes of protein expression in different cell states. ASAPRatio features a statistical approach which is valid if the expression level of a large number of identified proteins does not change between the two cell states:

A distribution of the logarithm (base=10) of all unique peptide ratios in an LC-ESI-MS/MS experiment is first generated:

- the dominant peak in the distribution is attributed to proteins of unchanged abundance and the ASAPRatio program fits the peak with a normal distribution (central limit theorem)
- The probability of the protein not changing in abundance is described statistically by the p value making data of large-scale protein profiling experiments comparable.

The result of this final step of the process is a calculated significance of abundance change for each identified protein.

Figure 4-14, from [72], summarizes ASAPRatio algorithm main steps.

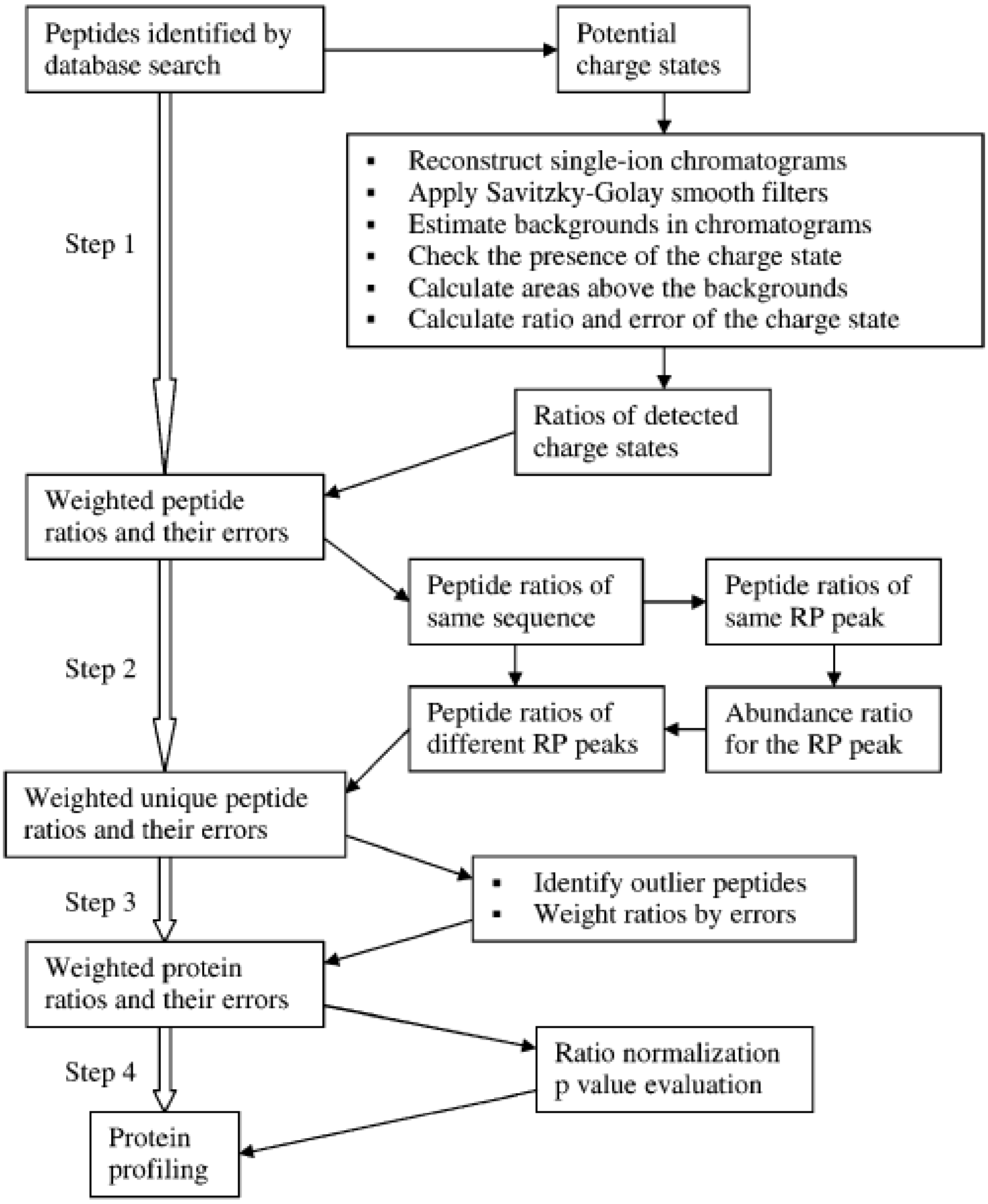

Peptides identified by database search
Potential charge states
▪ Reconstruct single-ion chromatograms
▪ Apply Savitzky-Golay smooth filters
▪ Estimate backgrounds in chromatograms
▪ Check the presence of the charge state
▪ Calculate areas above the backgrounds
▪ Calculate ratio and error of the charge state
Step 1
Ratios of detected charge states
Weighted peptide ratios and their errors
Peptide ratios of same sequence
Peptide ratios of same RP peak
Step 2
Peptide ratios of different RP peaks
Abundance ratio for the RP peak
Weighted unique peptide ratios and their errors
Step 3
▪ Identify outlier peptides
▪ Weight ratios by errors
Weighted protein ratios and their errors
Step 4
Ratio normalization p value evaluation
Protein profiling

FIGURE 4-14 FLOWCHART OF ASAPRATIO PROCEDURE TO DETERMINE PROTEIN QUANTIFICATION.

4.2.1.2 MASPECTRAS

MAss SPECTRometry Analysis System (MASPECTRAS) [66,67] is a platform for management and analysis of proteomics LC-MS/MS data. MASPECTRAS is based on the Proteome Experimental Data Repository (PEDRo) [50] relational database schema and follows the guidelines of the Proteomics Standards Initiative (PSI). This is a web-based platform with a back-end database and it relies on the Java 2 Enterprise Edition development platform. The platform is scalable and enables the outsourcing of computationally intensive tasks to a computing cluster. The data model captures data concerning experimental design and at all other subsequent steps leading up to evaluation and result export (see Figure 4-15). The MASPECTRAS imports and parses search results from SEQUEST [78], Mascot [79], Spectrum Mill, X! Tandem, and OMSSA and accepts mzXML and most instrument data formats. The capability to import and parse data from five search engines makes the platform universal and independent of the workflow performed by the proteomics research group.

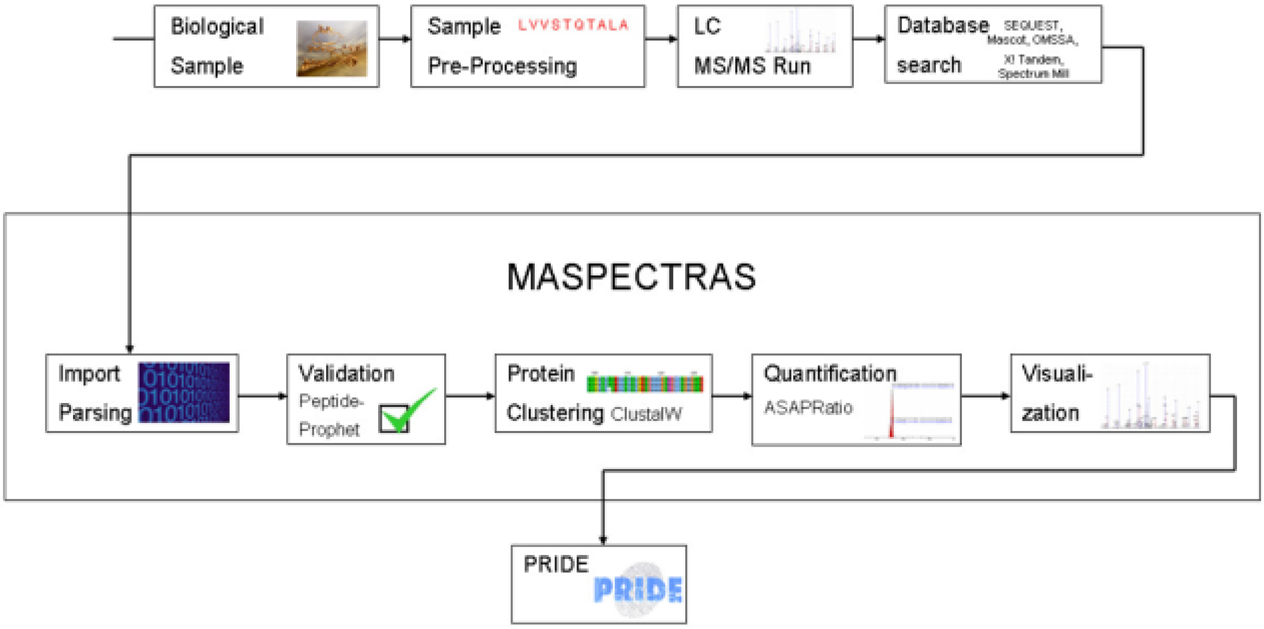

Biological Sample
Sample LVVSTQTALA Pre-Processing
LC MS/MS Run
Database search SEQUEST, Mascot, OMSSA, X! Tandem, Spectrum Mill
MASPECTRAS
Import Parsing
Validation Peptide-Prophet
Protein Clustering ClustalW
Quantification ASAPRatio
Visuali-zation
PRIDE
PRIDE

FIGURE 4-15 SCHEMATIC OVERVIEW OF THE ANALYSIS PIPELINE OF MASPECTRAS.

The system is not confined to a specific manufacturer and can therefore be used in labs equipped with different instruments. Moreover, MASPECTRAS is a system that provides the basis for consensus scoring between MS/ MS search algorithms. Peptides are validated using PeptideProphet [80] and the corresponding proteins clustered based on Markov clustering and multiple alignments. Then the peptides are quantified by the ASAPRatio algorithm, and the results stored in the database and exported to the public repository PRIDE [51]. Here below the implementation of ASAPRatio embedded in MASPECTRAS will be thoroughly described, since it will be used in the following of this thesis to validate the quantification performance of a newly proposed method.

To gain quantitative information the raw data from the mass spectrometer must be analyzed. The virtual chromatograms are calculated from the raw data; these are then smoothed and afterwards used to calculate the peak area. In order to be able to implement improvements of the ASAPRatio algorithm it was reprogrammed for the Java programming language. In MASPECTRAS implementation the m/z range for the chromatogram is user-definable. The chromatogram of one charge state is calculated by the summation of the ion intensities, smoothed tenfold by repeated application of the Savitzky-Golay smooth filtering method. For each isotopic peak, center and width are determined. The peak width is primarily calculated by using the standard ASAPRatio algorithm and for further peak evaluation an additional algorithm for recognizing peaks with saddle points has been implemented. With this algorithm, a valley (a local minimum of the smoothed signal) is recognized to be part of the peak and added to the area. The calculated peak area is determined as the average of the smoothed and the unsmoothed peak. Background noise, which is estimated from the average signal amplitude of the peak's neighborhood (50 chromatogram value pairs above and below the respective peak's borders), is subtracted from this value. The peak error is estimated as the difference between the smoothed and the unsmoothed peaks. A calculated peak area is accepted when the calculated peak area is bigger than the estimated error and the peak value is at least twice the estimated background noise. The peak area is otherwise set to zero. The calculation takes place automatically in the course of the analysis pipeline of MASPECTRAS. The identified peptides are combined into groups (peptides having the same sequence and same modification). These groups are then further subdivided according to their charge state. For each subgroup the median over the masses of the found peptides is calculated. For the calculation of the chromatogram this median is taken as the center of the m/z range, and not the in silico calculated ideal value. The reason for this approach is that the results generated by mass spectrometer are subject to variable error that is dependent on the instrument that is used. Normally, the error in m/z direction remains more or less constant for a given peptide. Despite this, median is chosen, because it allows more robust identification of outliers and false positives. The calculation can take place in MASPECTRAS directly or on a computing cluster, according to the number of peptides requiring quantitation. The threshold for job delegation can be set in a configuration file. A threshold is useful because the transfer of big MS raw data files is time-consuming and not feasible for a small number of peptides. Starting with approximately 50 peptides the gain in time increases almost in linear proportion to the number of processors used. After the calculation is finished, the retrieved peak areas are assigned to the peptides in the database and permanently stored. This module has been implemented as an adduct to the rest of the pipeline. The data can be analyzed by the user during the quantification process. To validate MASPECTRAS quantification performance, the quantitative analysis was performed with MSQuant [81], PepQuan (Bioworks 3.2 – Thermo Electron), and ASAPRatio as implemented in MASPECTRAS. The system provides customizable data retrieval and visualization tools, as well as export to PRoteomics IDEntifications public repository (PRIDE). The integration of peptide validation, protein grouping and quantification algorithms in conjunction with visualization tools is important for the usability and acceptability of the system. Particularly the inclusion of a quantification algorithm in the pipeline is of interest since more and more quantitative studies are initiated. The results of MASPCTERAS validation experiment showed that the performance of ASAPRatio was superior to MSQuant and PepQuan. The MASPECTRAS platform offers researchers an environment for the rapid analysis of large-scale proteomics experiments. Due to its modular design it is flexible enough to easily accommodate future changes in proteomics data management.

4.2.1.3 MSQUANT

MSQuant [82] quantifies data generated from Applied Biosystems/MDS-Sciex, Thermo Fisher Scientific, Micromass/Waters and it is compliant to all labeling techniques and to label-free, too. MSQuant allows integration of data from advanced acquisition schemes and optimal use of the raw data resulting in very high quality identifications and quantitation. Its main modules can be seen in Figure 4-16 where the published flowchart is reported [82].

MSQuant quantifies stable isotopically labeled pairs or triplets on the basis of peptide identifications (rather than directly from the data) and requires at least one of the members of a SIL pair to be identified by MS/MS, which is then used to calculate the position for all other partner peaks. MSQuant uses an algorithm that centers the quantitation on the actual peaks. This is important for Finnigan LTQ-FT data where the masses in some MS spectra are shifted due to space-charge effects. Thus it is not necessary to widen the quantitation mass window to account for this effect and which would introduce the risk of affecting the quantitation result with data from unrelated peaks. Users can click on the result for any MS scan and view the corresponding raw data. Often, single scans are unreliable due to interference from co-eluting peaks, for instance. These scans can be removed from consideration under user-control. No quantitation assessment was provided.

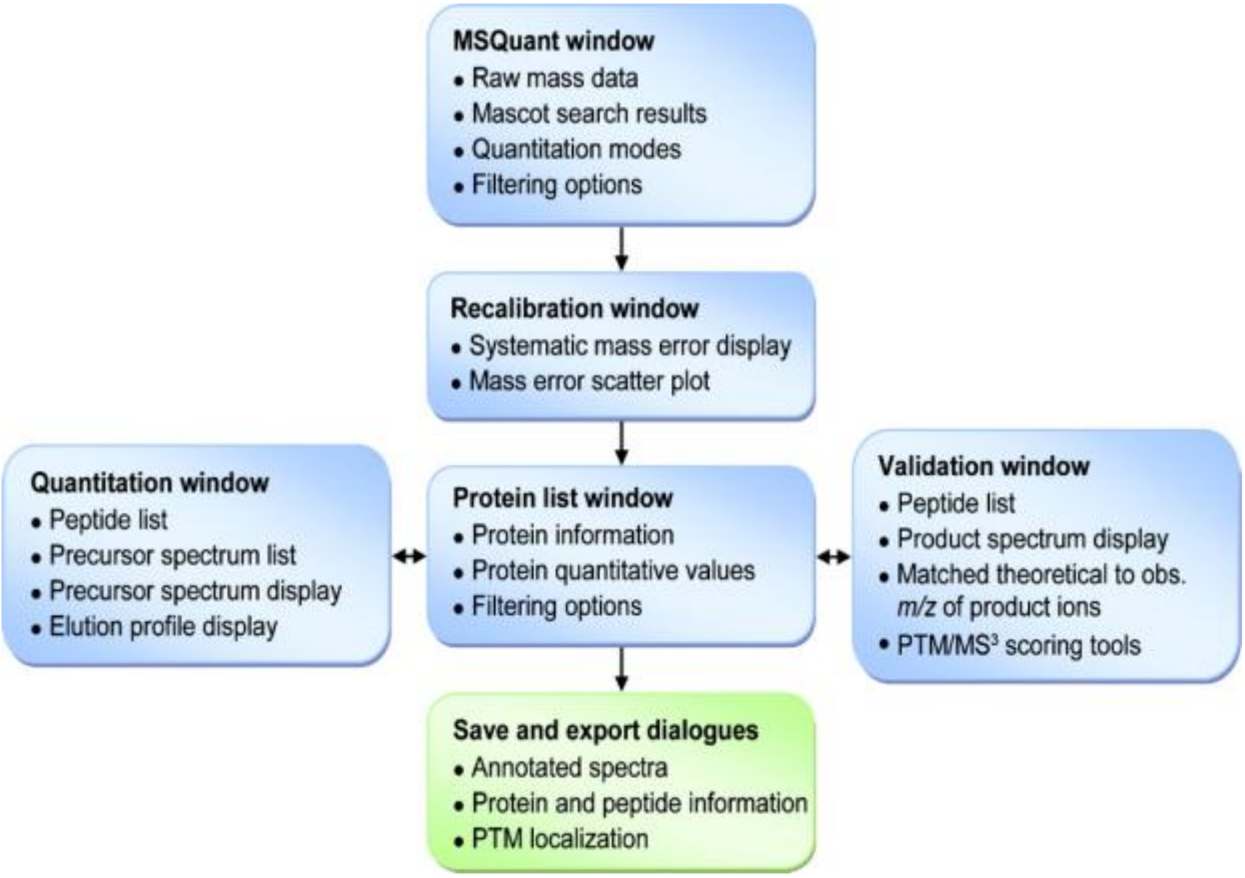
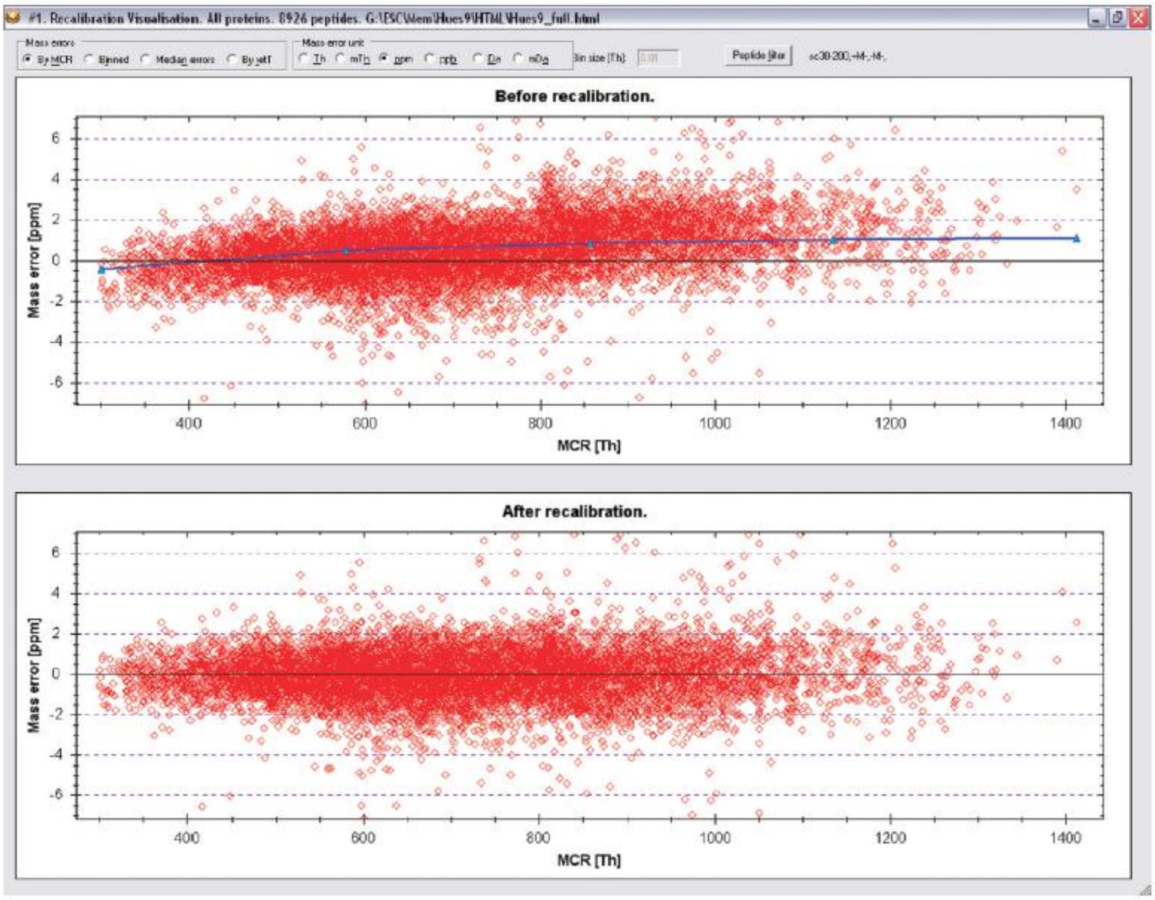

MSQuant window
• Raw mass data
• Mascot search results
• Quantitation modes
• Filtering options
Recalibration window
• Systematic mass error display
• Mass error scatter plot
Quantitation window
• Peptide list
• Precursor spectrum list
• Precursor spectrum display
• Elution profile display
Protein list window
• Protein information
• Protein quantitative values
• Filtering options
Validation window
• Peptide list
• Product spectrum display
• Matched theoretical to obs. m/z of product ions
• PTM/MS³ scoring tools
Save and export dialogues
• Annotated spectra
• Protein and peptide information
• PTM localization

FIGURE 4-16 MAIN APPLICATION WINDOWS OF MSQUANT. THE START SCREEN ASSOCIATES MASCOT RESULT FILES WITH THE CORRESPONDING RAW DATA FILES AND SPECIFIES PARAMETERS AND FILTERS FOR PARSING THE MASCOT FILE INTO MSQUANT. THE RECALIBRATION WINDOW ALLOWS THE USER TO EVALUATE PEPTIDE MASS ACCURACY BEFORE AND AFTER RECALIBRATION. THE PROTEIN LIST WINDOW IS THE MAIN DOCUMENT WINDOW AND CONTAINS A LIST OF IDENTIFIED PROTEINS. THIS WINDOW INTERFACES WITH MODULES FOR THE ANALYSIS OF SEQUENCE AND QUANTITATIVE INFORMATION EXTRACTED FROM THE PRECURSOR ION AND PRODUCT ION SPECTRA, RESPECTIVELY. MSQUANT STORES ALL DATA FOR AN EXPERIMENT IN A DOCUMENT FILE AND EXPORT ANNOTATED SPECTRA AND DATA IN VARIOUS REPORT FORMATS.

MSQuant main characteristic is the iterative recalibration (see Figure 4-17, from [82]) which improves the mass accuracy of the instrument: the observed vs calculated masses of high scoring peptides are used as internal calibrants. Optimal instrument-dependent calibration constants are calculated from the observed versus calculated masses of these peptides and these are then applied to all measured masses. The overall improvement in average mass accuracy is visualized in a separate window with various display options that provide the user with an immediate evaluation of the data quality and, thus, instrument performance and optimal database search parameters. A script changes the precursor masses in the peak list file after which a second search can be performed using the improved mass tolerance. MSQuant developers claim that this simple algorithm improves the mass accuracy of the instrument several fold, leading to much more specific search results.



#1. Recalibration Visualisation. All proteins. 8926 peptides. G:\ESC\dem\Hues9\HTML\Hues9_full.html

Mass errors: ● By MCR ○ Binned ○ Median errors ○ By retT

Mass error unit: ○ Th ○ mTh ● ppm ○ ppb ○ Da ○ mDa

Bin size [Th]: 0.01

sc30-200,+M-,-M-,

FIGURE 4-17 SCREENSHOT OF THE RECALIBRATION WINDOW IN MSQUANT. THIS WINDOW VISUALIZES THE PEPTIDE MASS ERRORS OF A DATA SET BEFORE AND AFTER RECALIBRATION. THE TREND LINE FOR THE 8926 HIGH SCORING PEPTIDES INDICATES A SMALL SYSTEMATIC CALIBRATION ERROR.

In the MSQuant framework, they first applied a Post Translational Modification (PTM) probability score (PTM-score) for MS3 experiments based on assigning a probability that the observed fragments match the fragments calculated for a given sequence by chance and then further developed the algorithm for phosphorylation matching. It iterates through all the possible modification sites and generates a score based on the number of supporting fragment masses, including handling the placement of several phosphorylation sites in a sequence, each of which may have different probabilities. While it was developed for phosphorylation, the principles underlying the PTM-score are of a general nature and can be used for any modification. MSQuant also allows evaluation of the PTM score by displaying the calculated fragment ions for any combination of the possible site-specific modifications for the MS/MS experiment as proposed by the scoring algorithm. Toggling between these possibilities gives valuable information about how much better the top scoring site localization is as compared to other interpretations.

4.2.1.4 MAXQUANT

MaxQuant automatically identifies several hundred thousand peptides per SILAC-proteome experiment and allows statistically robust identification and quantification of more than 4,000 proteins in mammalian cell lysates. It embodies a search engine (Andromeda), whose identification efficiency has never been assessed from developers, the same holds for the quantification method. It quantifies SILAC or label free data only from Thermo Fisher Scientific High Res FT-Data (Orbitrap data). Output results are statistically analyzed by the embedded module Perseus (workflow in Figure 4-18).

The data analyzed in this thesis are low resolution data, therefore MaxQuant couldn’t be applied to our dataset. Since it is one of the most used quantitation software nowadays we briefly illustrate its algorithm focusing only on the quantitative part.

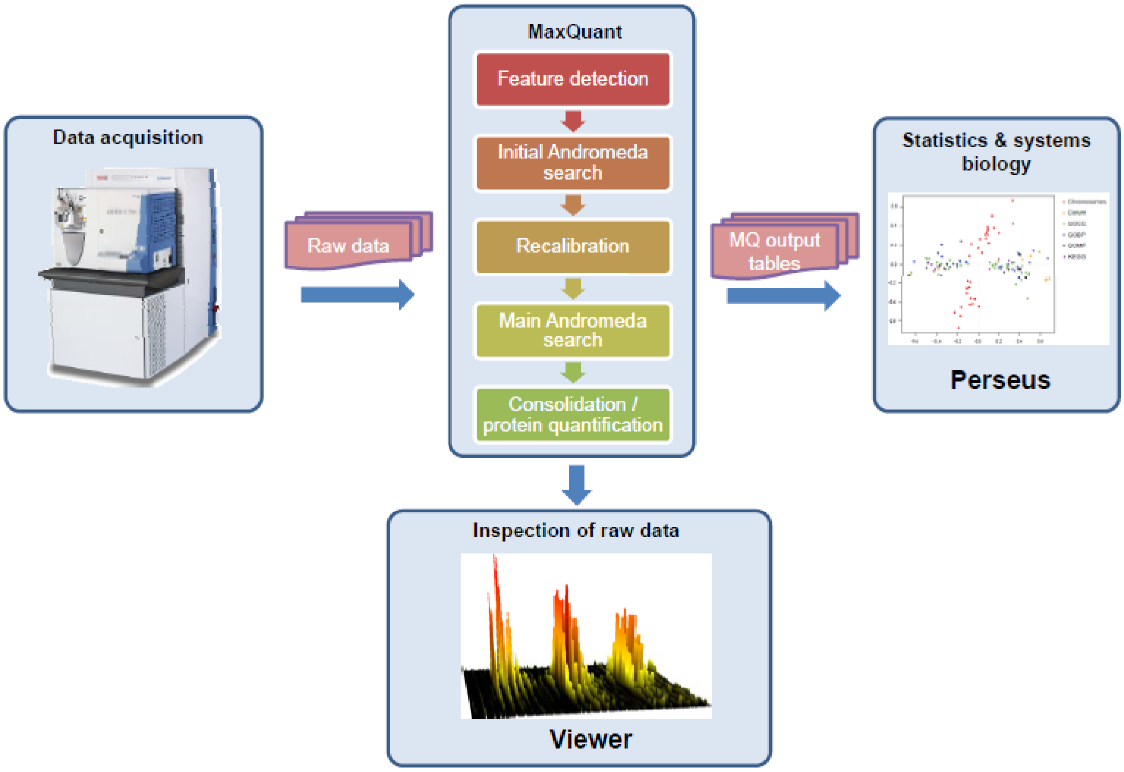


FIGURE 4-18 FLOWCHART ILLUSTRATING THE WORFLOW OF THE CURRENT MAXQUANT RELEASE.

To detect heavy-light SILAC partners MaxQuant considers all possible pairs of isotope patterns. Potential SILAC pairs are first required to have sufficient intensity correlation over elution time (allowing for some retention-time shift due to isotope effects) and to have equal charges. By default MaxQuant assumes at most three labeled amino acids per peptide. In order to get quantitation, for all possible cases, MaxQuant convolutes the two measured isotope patterns with the theoretical isotope patterns of the difference atoms, that is, the atoms that have to be added so that both peptides would have the same atomic composition. If the mass differences are within a bootstrap error computed in a previous step and if there is sufficient intensity correlation of the two isotope patterns in m/z dimension, the peaks are associated as a SILAC pair. The resulting isotope patterns should only differ by a global factor which is the ratio between the heavy and light peptide. To determine this ratio all corresponding 2D centroid intensities are paired. To these intensity pairs a straight line through the origin is fitted, whose slope is the desired ratio. The linear fit is done in a robust way, taking the least squares solution as initial value and then solving the best median fit equation iteratively by bisection.

In each LC-MS run, MaxQuant normalizes peptide ratios so that the median of their logarithms is zero, which corrects for unequal protein loading, assuming that the majority of proteins show no differential regulation (see Figure 4-19). Protein ratios are calculated as the median of all SILAC peptide ratios, minimizing the effect of outliers. MaxQuant normalizes the protein ratios to correct for unequal protein amounts.

MaxQuant finally calculates an outlier significance score for log protein ratios. As a P-value for detection of significant outlier ratios significance A is defined, which is the probability of obtaining a log-ratio of at least this magnitude under the null hypothesis that the distribution of log-ratios has normal upper and lower tails. For highly abundant proteins the statistical spread of unregulated proteins is much more focused than for low abundance ones. To capture this effect, another quantity, significance B is defined, which is calculated only on the protein subsets obtained by intensity binning. Bins of equal occupancy are defined.

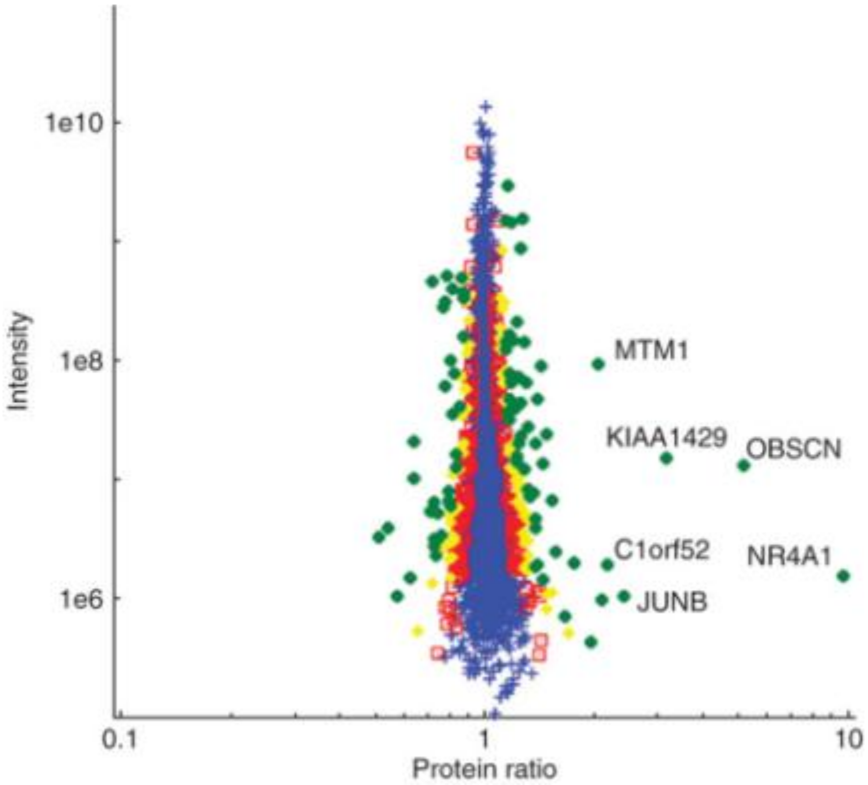


FIGURE 4-19 NORMALIZED PROTEIN RATIOS ARE PLOTTED AGAINST SUMMED PEPTIDE INTENSITIES. THE DATA POINTS ARE COLORED BY THEIR SIGNIFICANCE, WITH BLUE CROSSES HAVING VALUES >0.05, RED SQUARES BETWEEN 0.05 AND 0.01, YELLOW DIAMONDS BETWEEN 0.01 AND 0.001 AND GREEN CIRCLES <0.001.

4.2.1.5 CENSUS

Census [83] is able to quantify from either MS1 or MS/MS as well as to perform quantitative analyses based on both spectral counting and a LC-MS peak area approach using chromatogram alignment. It supports the following labeling strategies besides label free: ^{15}N, SILAC, iTRAQ. It accepts these data formats: DTASelect, mzXML and pepXML.

It can't be applied to the data analyzed during this PhD thesis work since it doesn't support ICPL labeled data, therefore we won't go into any detail. In Figure 4-20 a schematics representing Census main steps on labeled and label free data (from [83]).

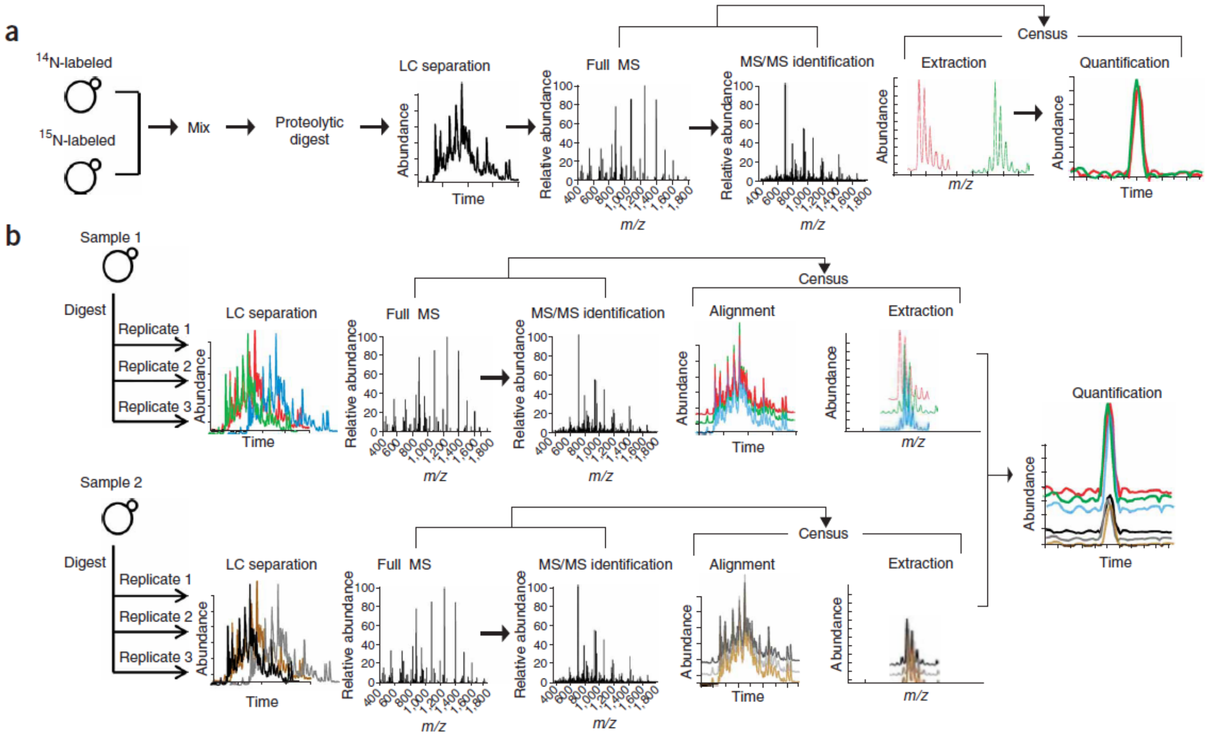


FIGURE 4-20 SCHEMATIC DETAILING THE QUANTITATIVE ANALYSIS CAPABILITIES OF CENSUS. (A) USE OF CENSUS WITH ISOTOPIC LABELING. (B) USE OF CENSUS WITH LABEL-FREE ANALYSIS. LC, LIQUID CHROMATOGRAPHY.

For isotopically labeled analyses, Census calculates peptide ion intensity ratios for each peptide pair using a linear least squares correlation to calculate the ratio (i.e., slope of the line) and closeness of fit (i.e., correlation coefficient) between the data points of the unlabeled and labeled ion chromatograms. To determine protein ratios, weighted means of peptide ratios were calculated.

In contrast to the approach used for isotopic labeling experiments, Census compares peak areas for peptides evaluated in an isotope free analysis. After the alignment, of multiple data files, Census evaluates all identified peptides by first taking the union of the search results from each individual file. Therefore, a peptide only needs to be identified in one file to be evaluated with respect to the entire dataset. The average peak area and variance for each peptide is calculated from technical replicates. Protein abundances are evaluated after outliers are removed using the average of peptide measurements.

They provided a quantification assessment comparing the expected and measured relative abundances of 4 technical replicates of a 10-protein mix dataset using Census. In Figure 4-21 published results are reported [83].

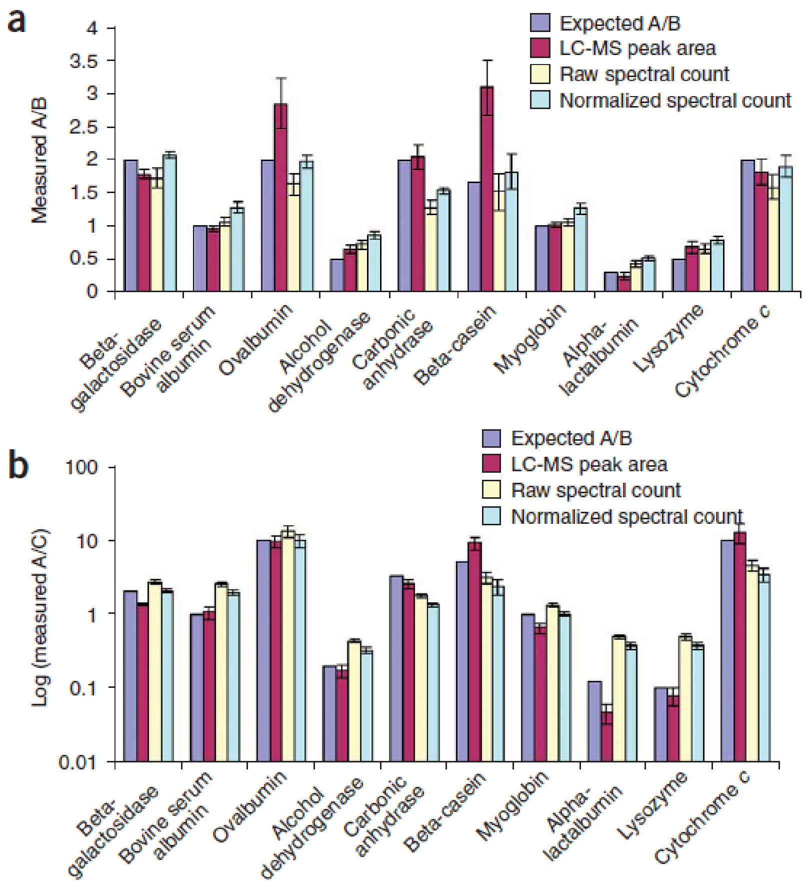


FIGURE 4-21 EXPECTED AND MEASURED RELATIVE ABUNDANCES OF TECHNICAL REPLICATES OF A 10-PROTEIN MIX DATASET USING CENSUS. (A) RATIO OF THE SIGNALS MEASURED FOR A MIXTURE OF SAMPLE A OVER SAMPLE B. (B) RATIO OF THE SIGNALS FOR A MIXTURE OF SAMPLE A OVER THAT OF SAMPLE C USING DIFFERENT STRATEGIES INCLUDING LC-MS PEAK AREAS, SPECTRAL COUNTING WITHOUT NORMALIZATION AND SPECTRAL COUNTING WITH NORMALIZATION. A TOTAL OF FOUR REPLICATE ANALYSES WERE PERFORMED FOR EACH MIXTURE AND VARIANCE WAS DETERMINED AS THE STANDARD DEVIATION.

4.2.1.6 OPENMS

OpenMS [84,85] is an open-source C++ library for LC/MS data management and analyses. It offers an infrastructure for the development of mass spectrometry related software. OpenMS is a free software available under the LGPL. OpenMS covers a wide range of functionalities needed to develop software for the analysis of high throughput protein separation and mass spectrometry related data: among others algorithms for signal processing, feature finding, visualization, map mapping and peptide identification (see Figure 4-22). OpenMS will be kept compatible with the upcoming Proteomics Standard Initiative (PSI) formats for MS data.

OpenMS has been successfully used for the implementation of The OpenMS Proteomics Pipeline (TOPP) [85]. TOPP is a set of computational tools that can be chained together to tailor problem-specific analysis pipelines for HPLC-MS data (see Figure 4-22). It transforms most of the OpenMS functionality into small command line tools that are the building blocks for more complex analysis pipelines.

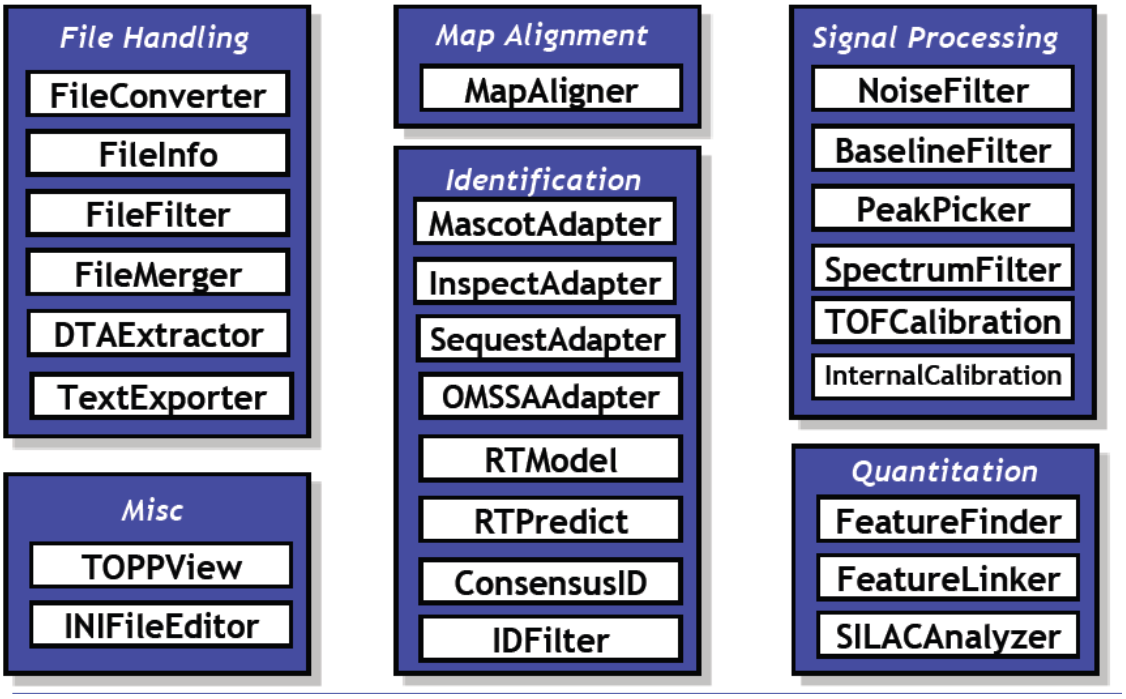


FIGURE 4-22 SOME OF THE MODULES IMPLEMENTED IN TOPP USING OPENMS.

OpenMS contains several algorithms for peptide quantitation based on model fitting [86,87] . Using the data structures provided by OpenMS and these algorithms, users are enabled to implement data analysis code for various complex quantitation tasks (labeled/unlabeled strategies, relative/ absolute quantitation). No quantitation assessment was provided, but the use of these algorithms improved quantitation accuracy in a complex absolute quantitation scenario (myoglobin in human blood plasma) while drastically reducing analysis times [88].

4.2.2 ANALYSIS ISSUES

At the state of art, LC-MS data analysis algorithms, especially for low resolution data, work on chromatographic 2D data. The chromatogram associated to a certain peptide is often extracted by integration of intensities in a defined m/z range (see Figure 4-23).

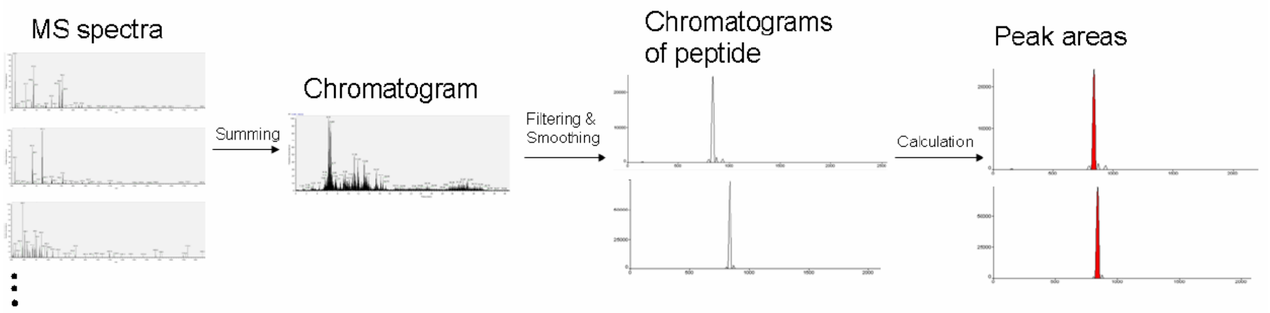
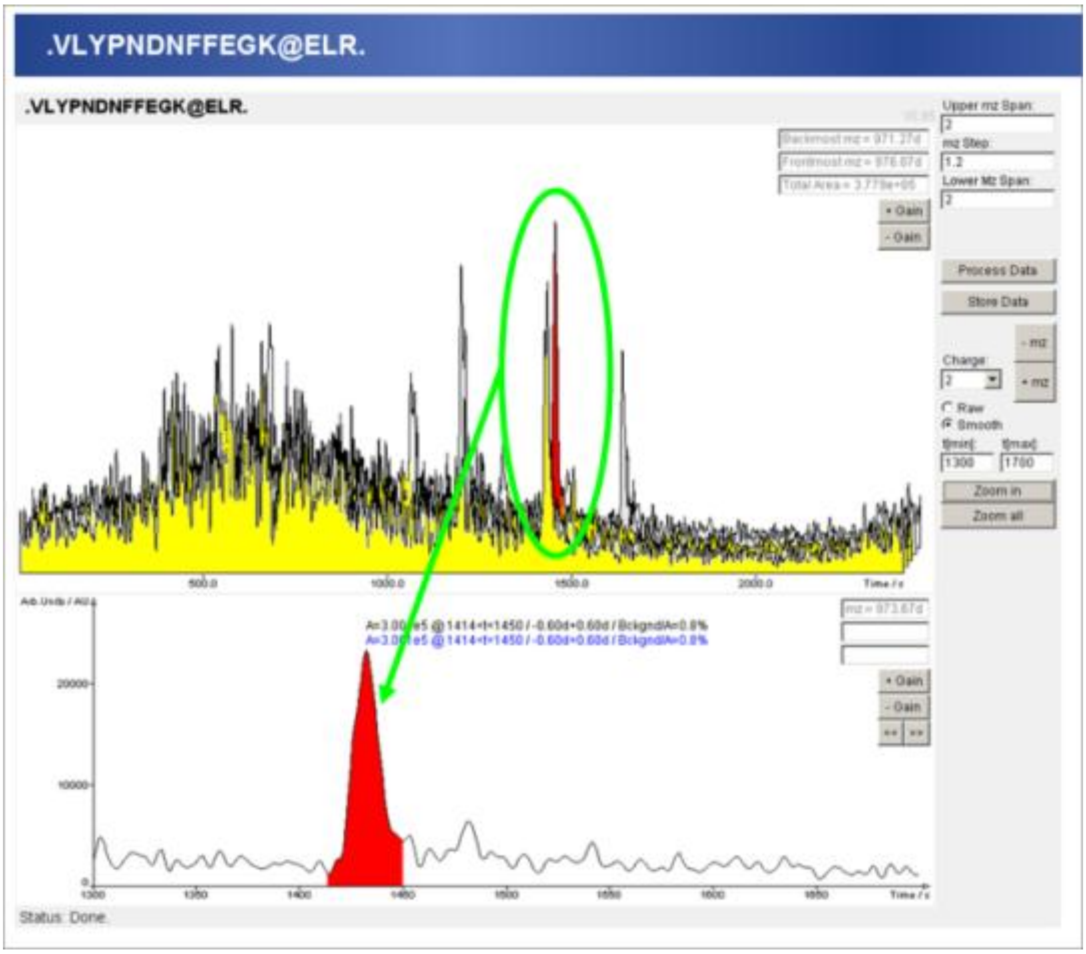


FIGURE 4-23 SCHEMATIC OVERVIEW OF THE RELATIVE QUANTIFICATION PROCESS. FOR THE CALCULATION OF A PEPTIDE ONLY THE MASS FLOW OF THE PEPTIDE IS OF INTEREST. THEREFORE CONTRIBUTIONS OF THE MASSES OF THE PEPTIDE ARE TAKEN INTO ACCOUNT. THE RESULTING CHROMATOGRAM IS SMOOTHED AFTERWARDS. DUE TO THE FACT THAT THE PEPTIDE CAN OCCUR AT DIFFERENT CHARGE STATES SEVERAL CHROMATOGRAMS HAVE TO BE TAKEN INTO CONSIDERATION. THE AREA BELOW THE CHROMATOGRAM CAN BE CALCULATED AS AN INDICATOR FOR THE AMOUNT OF PEPTIDE WHICH ENTERED THE MASS SPECTROMETER.

Such an approach, reducing a 3D signal to a 2D signal does not involve just a complexity reduction, but the loss of the LC-MS instrumentation resolving power. Thus, meaningful information is wasted, causing neighboring peaks to overlap along time dimension, resulting in unreliable quantifications (see Figure 4-24).



.VLYPNDNFFEGK@ELR.

.VLYPNDNFFEGK@ELR.

Backmost mz = 971.27d

Frontmost mz = 976.07d

Total Area = 3.779e+05

Upper mz Span: 2

mz Step: 1.2

Lower Mz Span: 2

Charge: 2

○ Raw

● Smooth

t[min]: 1300 t[max]: 1700

mz = 973.67d

A=3.00 e5 @ 1414<t<1450 / -0.60d+0.60d / BckgndA=0.8%

A=3.00 e5 @ 1414<t<1450 / -0.60d+0.60d / BckgndA=0.8%

Status: Done

FIGURE 4-24 PEPTIDES OVERLAPPING ON THE RETENTION TIME DIMENSION (PANEL ABOVE, RED AND YELLOW) MERGED TOGETHER (PANEL BELOW, IN RED) AFTER THE INTEGRATION ALONG THE M/Z DIMENSION AND THE SMOOTHING OPERATED BY THE PROCESSING ALGORITHMS.

For this reason a 3D approach reliably defining the borders of each peak is required. In fact, the 2D-LC-MS technique effectively separates peptides in the m/z and time dimensions. As a result, raising resolving power, LC-MS minimizes the overlapping of signals associated to peptides having similar electrochemical properties. Moreover, the profile acquisition mode enhances signal information content. Therefore, in this PhD research project we tried to exploit both data features to improve the quantification.

5 DATASET

In this chapter we describe the dataset used to evaluate performance of both the data handling solution and the quantification algorithm.

It consists of a controlled mixture of ICPL-labeled proteins (bovine serum albumin (UniprotKB: P02769), human apotransferrin (UniprotKB: P02787) and rabbit phosphorylase b (UniprotKB: P00489)). They were mixed at seven different light to heavy ratios (1:1, 1:2, 1:5, 1:10, 2:1, 5:1, 10:1) in triplicates.

The great advantage of so structured datasets relies in the fact that they enable to perform a reliable performance assessment. Very few studies have been published so far regarding the validation of algorithms for quantitative MS-based proteomics. Making use of these data we can test and compare several quantification algorithms.

Data were produced by the staff of the protein chemistry facility at the Research Institute of Molecular Pathology, Vienna. We were provided with these data by the Institute for Genomics and Bioinformatics and Christian-Doppler Laboratory for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria. Data are publicly available from MASPECTRAS[66] web site following the directions given in:

https://maspectras.genome.tugraz.at/maspectras/FileProvider?type=publicDownload&fileName =MASPECTRASPublishedDataHelp.pdf.

5.1 MATERIALS

Proteins were purchased from Sigma as lyophylized, dry powder. Solvents (HPLC grade) and chemicals (highest available grade) were purchased from Sigma, TFA (trifluoroacetic acid) was from Pierce. The ICPL (isotope coded protein label) chemicals kit was from Serva Electrophoresis and this kit contained reduction solution with TCEP (Tris (2-carboxy-ethyl) phosphine hydrochloride), cysteine blocking solution with IAA (Iodoacetamide), stop solutions I and II and the labeling reagent nicotinic acid N-hydroxysuccinimide ester as light (6 12C in the nicotinic acid) and heavy (6 13 C) form as solutions. Trypsin was purchased from Sigma at proteomics grade.

5.2 ICPL LABELING OF PROTEINS

Proteins were dissolved with TEAB (Tetraammonium bicarbonate) buffer (125 mM, pH 7.8) in three vials to a final concentration of 5 mg/ml each. A 40 µl aliquot was used for reduction of disulfide bonds between cysteine side-chains and blocking of free cysteines. For reduction of disulfide bonds 4 µl of reduction solution were added to the aliquot and the reaction was carried out for 35 min at 60°C. After cooling samples to room temperature, 4 µl of cysteine blocking solution were added and the samples were sat in a dark cupboard for 35 min. To remove excess of blocking reagent 4 µl of stop solution I were added and samples were put on a shaker for 20 minutes. Protein aliquots were split to two samples which contained 20 µl each. First row of samples was labeled with the 12C isotope by adding 3 µl of the nicotinic acid solution which contained the light reagent. Second row was labeled with the heavy reagent and labeling reaction was carried out for 2 h and 30 min while shaking at room temperature.

5.3 PROTEOLYTIC DIGESTION OF PROTEINS

Protein solutions were diluted using 50 mM NH4HCO3 solution to a final volume of 90 µl. 10 µl of a fresh prepared trypsin solution (2.5 µg/µl) were added and the proteolysis was carried out at 37°C over night in an incubator. The reaction was stopped by adding 10µl of 10% TFA. The peptide solutions were diluted with 0.1 % TFA to give 1 nM final concentration. From these stock solutions samples for MS/MS analysis which contained defined ratios of heavy and light were made up by mixing the solutions of light and heavy labeled peptides.

5.4 HPLC AND MASS SPECTROMETRY

To separate peptide mixtures prior to MS analysis, nano reverse phase high-performance liquid chromatography (nanoRP-HPLC) was applied on the Ultimate 2 Dual Gradient HPLC system (Dionex, buffer A: 5% acetonitrile (ACN), 0.1% TFA, buffer B: 80% ACN, 0.1% TFA) on a PepMap separation column (Dionex, C18, 150 mm × 75 µm × 3 µm, 300 A). 500 fM of each mixture was separated three times using the same trapping and separation column to reduce the quantification error which comes from HPLC and mass spectrometry. A gradient from 0% B to 50% B in 48 min was applied for the separation; peptides were detected at 214 and 280 nm in the UV detector. The exit of the HPLC was online coupled to the electrospray source of the LTQ mass spectrometer (Thermo Electron). Samples were analyzed in centroid mode first to test digest and labeling quality. For the quantitative analysis the LTQ was operated in enhanced profile mode for survey scans to gain higher mass accuracy. Samples were mass spectrometrically analyzed using a top one method, in which the most abundant signal of the MS survey scan was fragmented in the subsequent MS/MS event in the ion trap. Although with this method a lower number of MS/MS spectra were acquired, the increased number of MS scans leads to a better determination of the eluting peaks and therefore provides improved quantification of peptides.

Data analysis was done with the Mascot Daemon (Matrix Science), BioWorks 3.2 (Thermo Electron) software packages using an in house database. To demonstrate the merging of results from search engines the ICPL labeled probes at a ratio of 1:1 were searched with Spectrum Mill A.03.02 (Agilent Technologies), X! Tandem (The Global Proteome Machine Organization) version 2006.04.01, and OMSSA 1.1.0 (NCBI).

6 DATA HANDLING: THE MZRTREE DATA STRUCTURE

In this chapter we present a novel data structure, called *mzRTree*, for the efficient handling of high-throughput LC-MS profile datasets. It combines a hybrid sparse/dense matrix representation of the data and a scalable index based on the R-tree [89] (see Figure 6-1). We show experimentally that mzRTree supports efficiently both 1D and 2D data accesses. In particular, mzRTree significantly outperforms other known structures used for LC-MS data on small and large peptide range queries, yielding in some cases orders of magnitude improvements. Furthermore, it still ensures best performance on the accesses for which the other data structures are optimized. The experiments also provide evidence that mzRTree is more space efficient, and exhibits good scalability on increasing dataset densities. In the following of this chapter the theoretical approach, its actual implementation and the performance validation will be comprehensively illustrated and finally discussed.

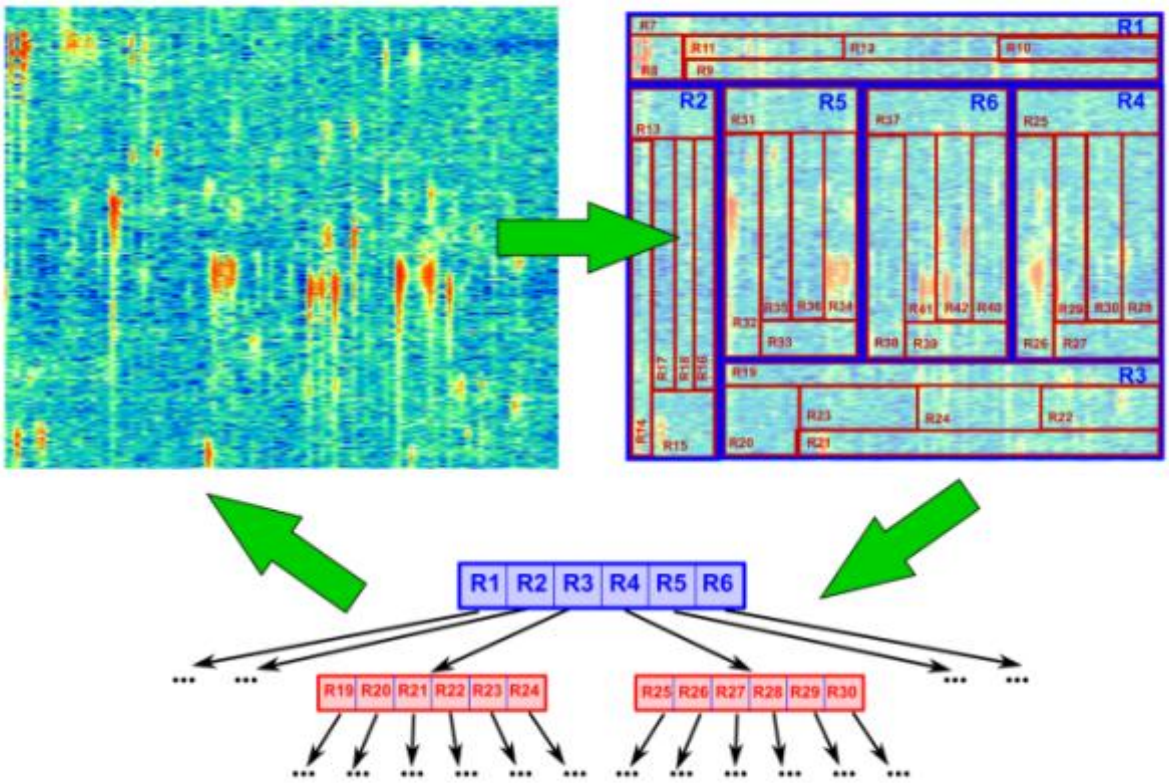


FIGURE 6-1 LC-MS DATA DIVIDED IN NESTED RECTANGLES AND INDEXED BY THE R-TREE. INDEXED RECTANGLES CAN BE EFFICIENTLY ACCESSED MAKING USE OF THE R-TREE.

Let us conceptually view an LC-MS dataset *D* as a matrix, where the rows are indexed by retention times, the columns by m/z values, and the entries are intensity values. A generic entry is denoted as (*rt*, *mz*; *I*), where *rt* and *mz* are the row and column indices, and *I* is the intensity value.

We store *D* using a hybrid sparse/dense matrix representation, as follows. First, we evenly subdivide the matrix into *K strips* of consecutive rows, where *K* is a user defined parameter. Then, each strip is in turn partitioned into a number of *bounding boxes* (*BBs*), each corresponding to a distinct range of m/z values. In our implementation, each BB corresponds to approximately 5 Da, and *K* is set in such a way to ensure that each strip fits in the main memory (*RAM*). A BB is characterized by four coordinates, namely: *top-rt* (resp., *bottom-rt*), which is the smallest (resp., largest) retention time of the BB's nonzero intensity entries; and *left-mz* (resp., *right-mz*), which is the smallest (resp., largest) m/z value of the BB's nonzero intensity entries. The BBs of a strip are stored consecutively in a file, and each strip is saved in a distinct file so that it can be efficiently loaded in the main memory during a range query. If half or more of the entries in a BB have nonzero intensity, then the BB is stored in the file using a dense matrix representation. Otherwise, a sparse representation is used storing the nonzero intensity entries in row-major order, indicating for each such entry the column (m/z value) and the intensity, and separating successive rows through special characters. In this fashion, each BB occupies a space proportional to the number of nonzero intensity entries it contains.

A *range query operation* on *D* takes as input two retention times rt_1, rt_2, and two m/z values, mz_1, mz_2, and returns all entries (*rt*, *mz*; *I*) in *D* such that $rt_1 < rt \leq rt_2$ and $mz_1 < mz \leq mz_2$. Accesses to chromatograms, spectra or peptide data can be easily expressed through range queries. In order to support efficient range query operations, we use an index implemented through a tree structure based on the *R-tree* [89], which is a well-known spatial data structure for managing geometric data.

Let *d* and *f* be two integer parameters, and let *G* be the set of nonempty BBs (i.e., BBs which contain at least one nonzero-intensity entry). Denote by *W* the cardinality of *G*. Our index consists of a balanced search tree whose leaves are associated with disjoint subsets of *G* forming a partition of *G*. The number of children of each internal node is proportional to *d* (the root, if internal, may have a smaller number of children) and each leaf is associated with a subset of size proportional to *f* of BBs in *G* (the root, if a leaf, may have less than *f* BBs). Each internal node of the tree is associated to the smallest submatrix of *D* which contains all BBs associated with its descendant leaves.

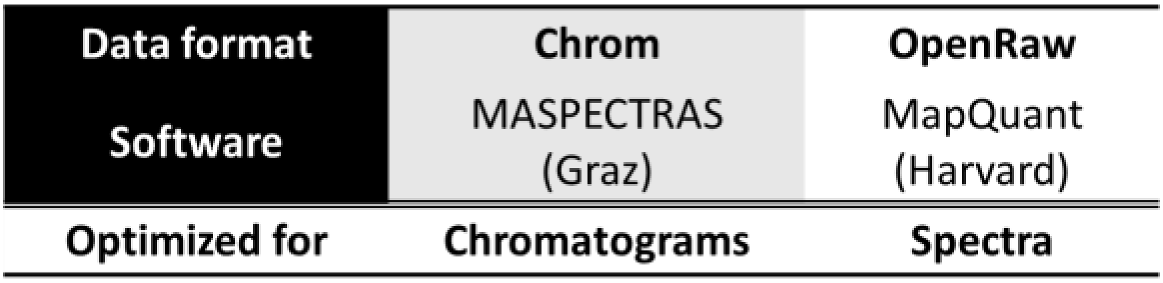
The execution of a range query requires to traverse all root-leaf paths ending in leaves associated with BBs that intersect the rectangle defined by the query, and to return all entries of interest. The complexity of a range query depends on the height of the tree, hence on the parameters *d* and *f*, and on the mapping of the BBs to the leaves. As for the choice of the partition parameters *d* and *f*, when dealing with massive datasets, which must be kept in secondary memory, it is convenient to impose that each node of the tree (except, possibly, the root) occupies a constant fraction of the minimal unit that can be transferred between the secondary memory and the RAM. Instead, for what concerns the mapping of the BBs to the leaves, several heuristics have been proposed in the literature (see [90] for relevant references).

In our implementation, we set d=6 and f=200, and the actual structure of the tree is recursively defined as follows, based on ideas in [89]. If $W \leq f$, the tree consists of a single leaf associated with the set G; otherwise, G is partitioned into six groups, G_i, for $1 \leq i \leq 6$, as follows. G_1 contains the $\lceil W/6 \rceil$ BBs with smallest top-rt; G_2 contains the $\lceil W/6 \rceil$ BBs in G- G_1 with smallest left-mz; G_3 contains the $\lceil W/6 \rceil$ BBs in G- G_1- G_2 with largest bottom-rt; G_4 contains the $\lceil W/6 \rceil$ BBs in G- G_1- G_2- G_3 with largest right-mz; G_5 contains the $\lceil W/6 \rceil$ BBs in G- G_1- G_2- G_3- G_4 with smallest left-mz; and G_6 contains the remaining BBs. The six groups are associated with the subtrees rooted at the children of the root, which are recursively organized in a similar fashion. Each leaf is thus associated with up to f=200 BBs, and it stores, for each of its BBs, the four coordinates (top-rt, bottom-rt, left-mz, right-mz) and a pointer to the file where the BB is stored together with the relative offset within the file. It can be easily shown that the height of the tree is proportional to $\log_6 (W/200)$.

We call *mzRTree* the whole data structure, which includes the actual data (i.e., the bounding boxes) stored in the files, and the tree index described above. We developed a Java implementation of mzRTree, which includes a method to build an mzRTree starting from an input dataset provided in mzXML/mzML format [59,61], and a method to perform a generic range query[1].

[1] The Java code implementing mzRTree is available for download at *http://www.dei.unipd.it/mzrtree*.

In this section, we describe how we evaluated mzRTree performance compared to Chrom and OpenRaw, which are two existing data structures used by Maspectras and MapQuant software packages and optimized for chromatograms and spectra based accesses, respectively (see Table 6-1). Specifically, we focused our analysis on the time required for a range query, the time required for building up the data structure, and the required hard disk space. Furthermore, we verified mzRTree scalability for what concerns range query times using datasets of increasing density, where the density of a dataset is defined as the ratio of the number of retention time and m/z value pairs associated with nonzero intensities to the overall number of retention time and m/z value pairs.

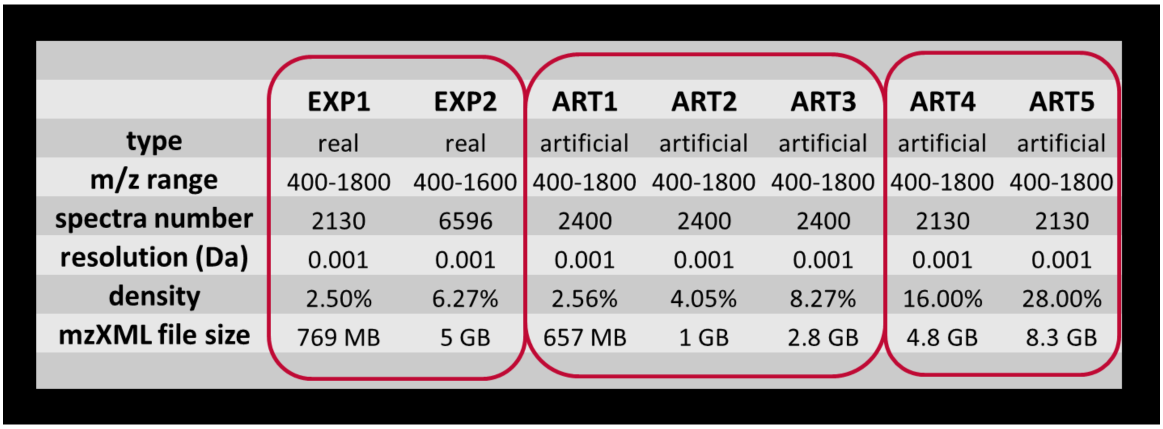


Data format	Chrom	OpenRaw
Software	MASPECTRAS (Graz)	MapQuant (Harvard)
Optimized for	Chromatograms	Spectra

TABLE 6-1 IT SUMMARIZES THE MAIN FEATURES OF THE DATA STRUCTURES USED IN THIS COMPARISON.

We compared mzRTree, Chrom and OpenRaw on seven LC-MS datasets, named EXP1, EXP2, ART1, ART2, ART3, ART4 and ART5, which are described below. The *EXP1* dataset consists of real profile data from a controlled mixture of ICPL-labeled proteins acquired in enhanced profile mode for survey scans to gain higher mass accuracy using a Finnigan LTQ linear ITMS (Thermo Electron) equipped with HPLC-NSI source. The *EXP2* is a real profile dataset acquired with a Waters ESI TOF Microchannel plate LCT Premier available on the PeptideAtlas public database. The *ART1*, *ART2* and *ART3* datasets have been generated by the LC-MS simulator LC-MSsim [91] using as input some peptide sequences from bovine serum albumin (UniprotKB: P02769), human apotransferrin (UniprotKB: P02787) and rabbit phosphorylase b (UniprotKB: P00489). Finally, the *ART4* and *ART5* datasets have been generated artificially by the following procedure: for each dataset, the user specifies some input parameters, namely, the number of spectra (i.e., the total number of retention times), the m/z range and resolution, and the density *d*; then, each spectrum is populated by assigning nonzero intensity values to positions corresponding to m/z values drawn from a uniform distribution until the density of the spectrum is *d*; clearly, if each spectrum has density *d*, then the final dataset will have density *d*. *ART4* and *ART5* are useful to evaluate the scalability of our data structure although they are not meaningful from a biological standpoint.

The characteristics of the aforementioned datasets are summarized in Table 6-2. Notice that the resolution shown in Table 6-2 is not the original data resolution (i.e., the instrumental resolution) but it is a suitable resolution, not smaller than the original one, which has been adopted in our data representation for uniformity with the other data representations used for comparison in the experiments. In particular, the Chrom files we used, adopted a 0.001 Da resolution: this resolution is higher than the maximum resolution achievable by the instruments used to acquire the experimental data. Therefore, our choice is conservative in the sense that it does not require any binning and, consequently, does not cause any loss of information.



	EXP1	**EXP2**	**ART1**	**ART2**	**ART3**	**ART4**	**ART5**
type	real	real	artificial	artificial	artificial	artificial	artificial
m/z range	400-1800	400-1600	400-1800	400-1800	400-1800	400-1800	400-1800
spectra number	2130	6596	2400	2400	2400	2130	2130
resolution (Da)	0.001	0.001	0.001	0.001	0.001	0.001	0.001
density	2.50%	6.27%	2.56%	4.05%	8.27%	16.00%	28.00%
mzXML file size	769 MB	5 GB	657 MB	1 GB	2.8 GB	4.8 GB	8.3 GB

TABLE 6-2 DATASETS' FEATURES. NOTICE THAT THE SPECTRA NUMBER IS REFERRED TO THE TOTAL NUMBER OF MS^1 SPECTRA AND RESOLUTION IS NOT THE INSTRUMENT RESOLUTION, AS EXPLAINED IN THE TEXT. RED CIRCLES ARE CLUSTERING SIMILAR DATASETS.

We compared mzRTree, Chrom and OpenRaw on four kinds of range queries: a rectangle covering all the retention times and a 5 Da range in the m/z dimension (*chromatograms*); a rectangle covering the entire m/z dimension and 20 retention times (*spectra)*; a rectangle of 5 Da and 60 retention times (*small peptide*); a rectangle of 5 Da and 200 retention times (*large peptide*).

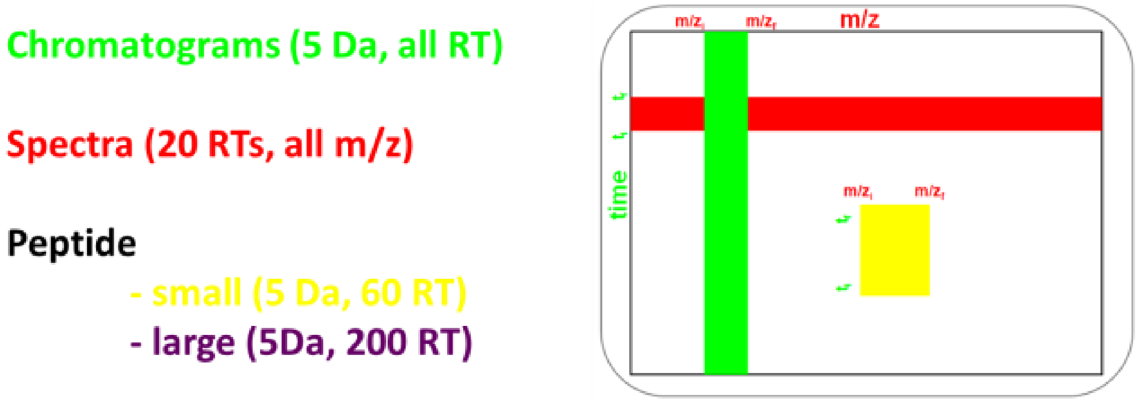


FIGURE 6-2 THE FIGURE VISUALLY ILLUSTRATES THE 3 MAIN KINDS OF DATA ACCESSES WE TESTED MZRTREE FOR.

We estimated the performance for each kind of range query summing the access times required to perform ten range queries spanning the whole dataset in order to avoid any local advantage. More precisely, we evaluated separately the time required for loading the internal variables used by each data structure every time it is invoked (*load time*) and the time actually needed to perform only the range query (*access time*). To reduce random variability, we computed both access and load times averaging over ten experimental repetitions. It is worth to notice that a spectra range query is more time consuming than a chromatograms range query, since the number of distinct m/z values is typically much bigger than the number of retention times.

Results on access times for the EXP1 and EXP2 datasets are shown in Figure 6-3 and Figure 6-4, respectively: mzRTree achieves the best performance on all kinds of range queries for both the smaller size and density dataset EXP1 and the larger size and density dataset EXP2.

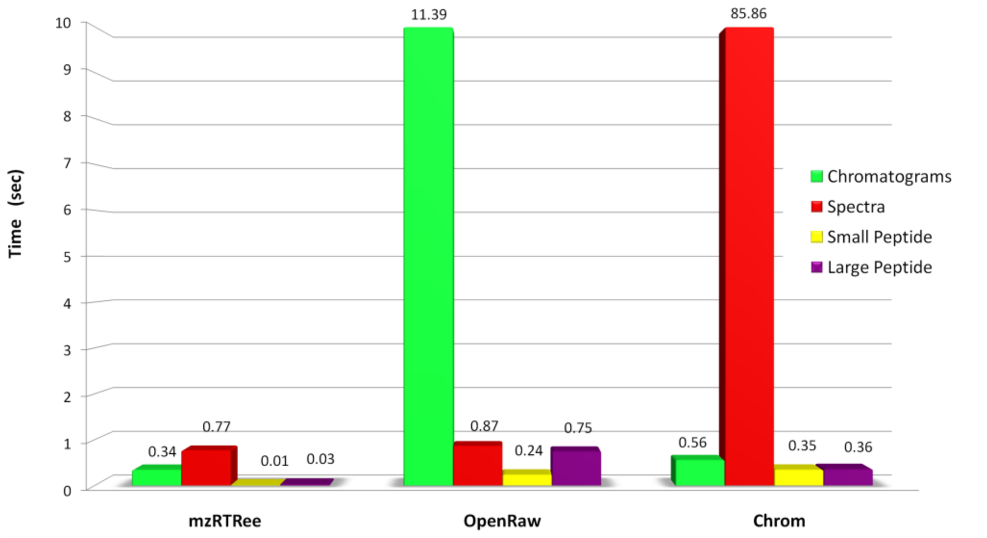


FIGURE 6-3 COMPARISON ON EXP1 DATASET AMONG MZRTREE, OPENRAW AND CHROM ON RANDOM CHROMATOGRAMS, SPECTRA AND SMALL/LARGE PEPTIDE RANGE QUERIES SPANNING THE WHOLE DATASET AS REGARDS ACCESS TIMES. EVERY COLORED BAR REFERS TO A DIFFERENT KIND OF RANGE QUERY. MZRTREE REACHES BEST PERFORMANCE ON ALL KIND OF RANGE QUERIES, OUTPERFORMING CHROM AND OPENRAW.

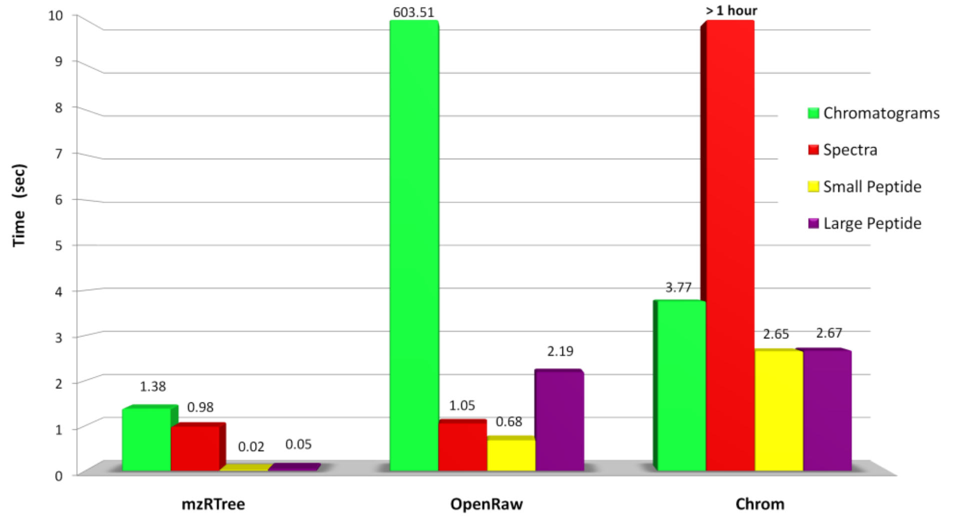


FIGURE 6-4 COMPARISON ON EXP2 DATASET AMONG MZRTREE, OPENRAW AND CHROM ON RANDOM CHROMATOGRAMS, SPECTRA AND SMALL/LARGE PEPTIDE RANGE QUERIES SPANNING THE WHOLE DATASET AS REGARDS ACCESS TIMES. EVERY COLORED BAR REFERS TO A DIFFERENT KIND OF RANGE QUERY. NOTICE HOW MZRTREE STILL REACHES BEST PERFORMANCE, OUTPERFORMING CHROM AND OPENRAW, ALSO ON THIS HIGHER DENSITY AND SIZE DATASET.

Furthermore, Figure 6-5 illustrates the access times for ten peptides in EXP1 using small and large peptide range queries, whose bounds refer to peptides actually identified by the Mascot search engine. mzRTree significantly outperforms Chrom and OpenRaw on small and large peptide range queries, and still ensures best performance on the accesses for which the other data structures are optimized, i.e., chromatograms for Chrom and spectra for OpenRaw.

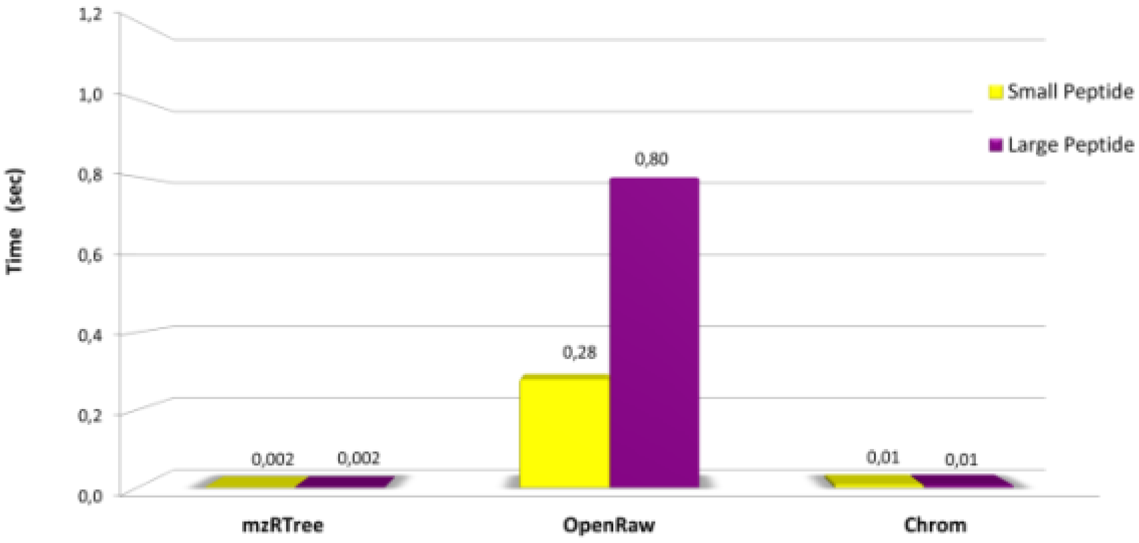


FIGURE 6-5 COMPARISON ON EXP1 DATASET AMONG MZRTREE, OPENRAW AND CHROM ON SMALL/LARGE PEPTIDE RANGE QUERIES RELATED TO MASCOT IDENTIFIED PEPTIDES AS REGARDS ACCESS TIMES: MZRTREE IS ONE ORDER OF MAGNITUDE FASTER THAN CHROM AND TWO ORDERS OF MAGNITUDE FASTER THAN OPENRAW.

The load time required by the three data structures is shown in Figure 6-6 for EXP1 and EXP2 datasets: we note that the load time is mainly independent of dataset features, and mzRTree still achieves the best performance. Since loading is required every time the data structures are invoked, it is convenient to perform many consecutive range queries in order to amortize its cost: the higher the load time, the more the range queries needed to amortize it.

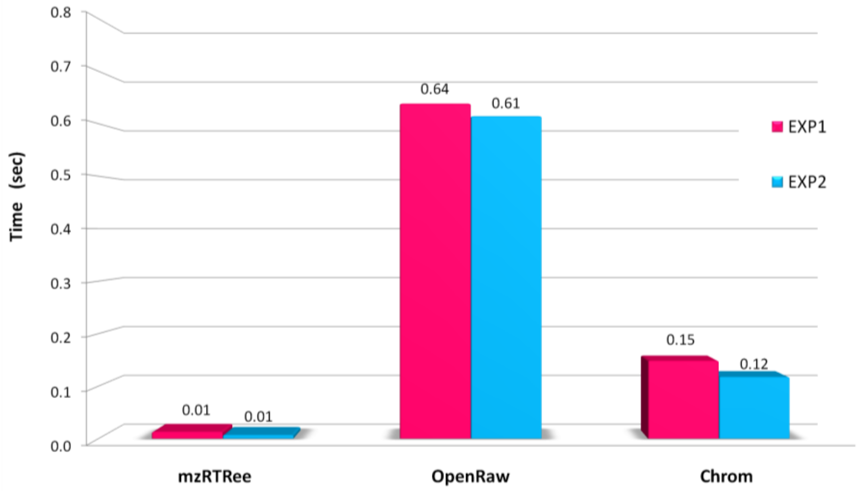


FIGURE 6-6 COMPARISON ON EXP1 AND EXP2 DATASETS AMONG MZRTREE, OPENRAW AND CHROM ON LOAD TIMES: MZRTREE IS ONE ORDER OF MAGNITUDE FASTER THAN CHROM AND OPENRAW. MZRTREE IS ONE ORDER OF MAGNITUDE FASTER THAN CHROM AND OPENRAW.

Even if the data structure creation takes place just once, we also estimated the creation time for mzRTree, Chrom and OpenRaw on EXP1. Notice that while mzRTree and Chrom creation starts from the mzXML file, the OpenRaw creation starts from the .RAW file, requiring the instrument vendor's software to be licensed and installed on the computer. We chose EXP1 because its size is small enough to fit in RAM, thus all three data structures evenly work at their best condition. As shown in Figure 6-7, mzRTree features an efficient creation time, even if OpenRaw reaches the best performance. However, notice that OpenRaw is advantaged since it starts from binary data instead of Base64 encoded data.

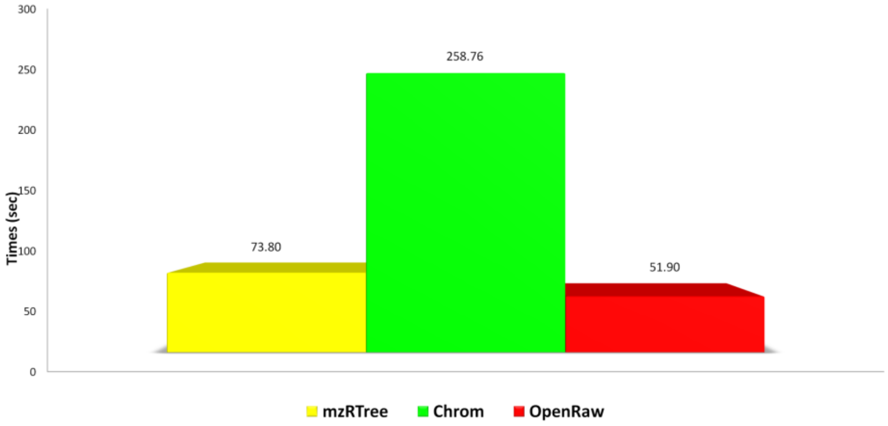


FIGURE 6-7 COMPARISON OF MZRTREE, CHROM AND OPENRAW AS REGARDS DATA STRUCTURES' CREATION TIME FOR EXP1 DATASET. WHILE MZRTREE AND CHROM CREATION STARTS FROM THE MZXML FILE, OPENRAW CREATION STARTS FROM THE .RAW FILE, REQUIRING THE INSTRUMENT VENDOR'S SOFTWARE TO BE LICENSED AND INSTALLED, HENCE IT STARTS FROM BINARY DATA INSTEAD OF BASE64 ENCODED DATA.

In Table 6-3 we provide the comparison of the space reduction using mzRTree, Chrom and OpenRaw compared to the mzXML hard disk space, which we chose as reference. mzRTree requires the smallest amount of space, hence it allows for cheaper storage and easier sharing of proteomics datasets. Besides, mzRTree storage requires at least 30% less hard disk space than XML based data formats, since mzRTree stores binary data instead of Base64 encoded data: it is a considerable amount of space saved, when taking into account RAID systems and backup systems. Observe that, since, for the sake of simplicity, we are ignoring MS level-two spectra, the space savings for the first two datasets are notably larger than 30%; however, this is not the case of the third dataset, which consists only of level-one spectra.

mzXML	***EXP1***	***EXP2***	***ART4***
mzRTree	53.71%	46.00%	25.00%
Chrom	37.84%	28.00%	-
OpenRaw	27.31%	18.00%	-10.42%

TABLE 6-3 SPACE REDUCTION REFERRED TO THE ORIGINAL MZXML FILE SIZE, CHOSEN AS REFERENCE. MZRTREE ALLOWS FOR A MORE EFFICIENT HARD DISK SPACE-SAVING STORAGE.

mzRTree can efficiently handle also tandem data; the user only needs to create the data structure for every MS/MS level of interest. Figure 6-8 shows that mzRTree provides efficient access times on tandem MS data for all kind of range queries, attaining for MS level 2 data the same performance as for MS level 1 data.

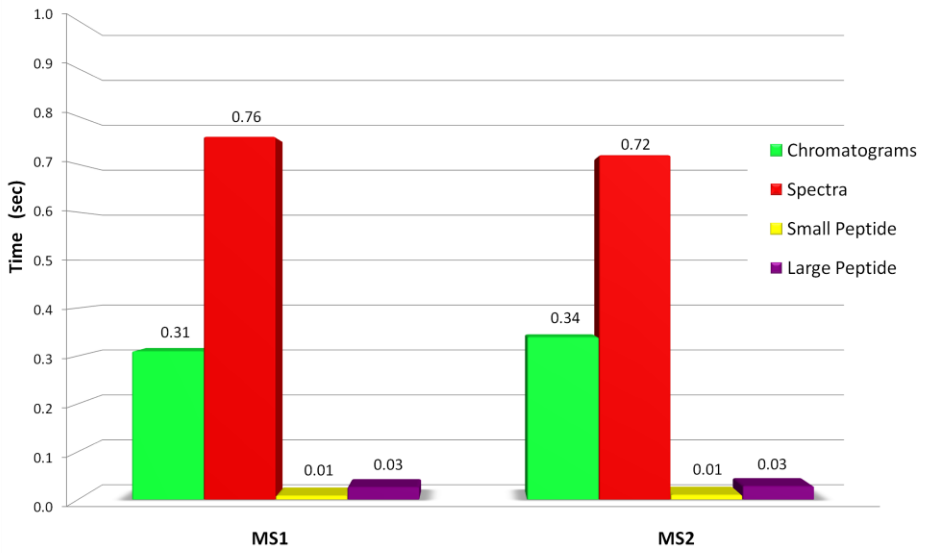


FIGURE 6-8 COMPARISON OF MZRTREE ACCESS TIMES ON MS^1 AND MS^2 LEVELS FOR EXP1 DATASET. THE PERFORMANCE OF MZRTREE IS INDEPENDENT OF THE MS LEVEL.

To test mzRTree scalability on increasing dataset densities and sizes we performed different range queries on the artificial datasets ART1, ART2, ART3, ART4 and ART5. Results are illustrated in Figure 6-9, which shows that mzRTree is fairly scalable as regards access and load time: as data density increases by a factor 10, the access time increases only by a factor 3 in the worst case, while the load time is almost constant.

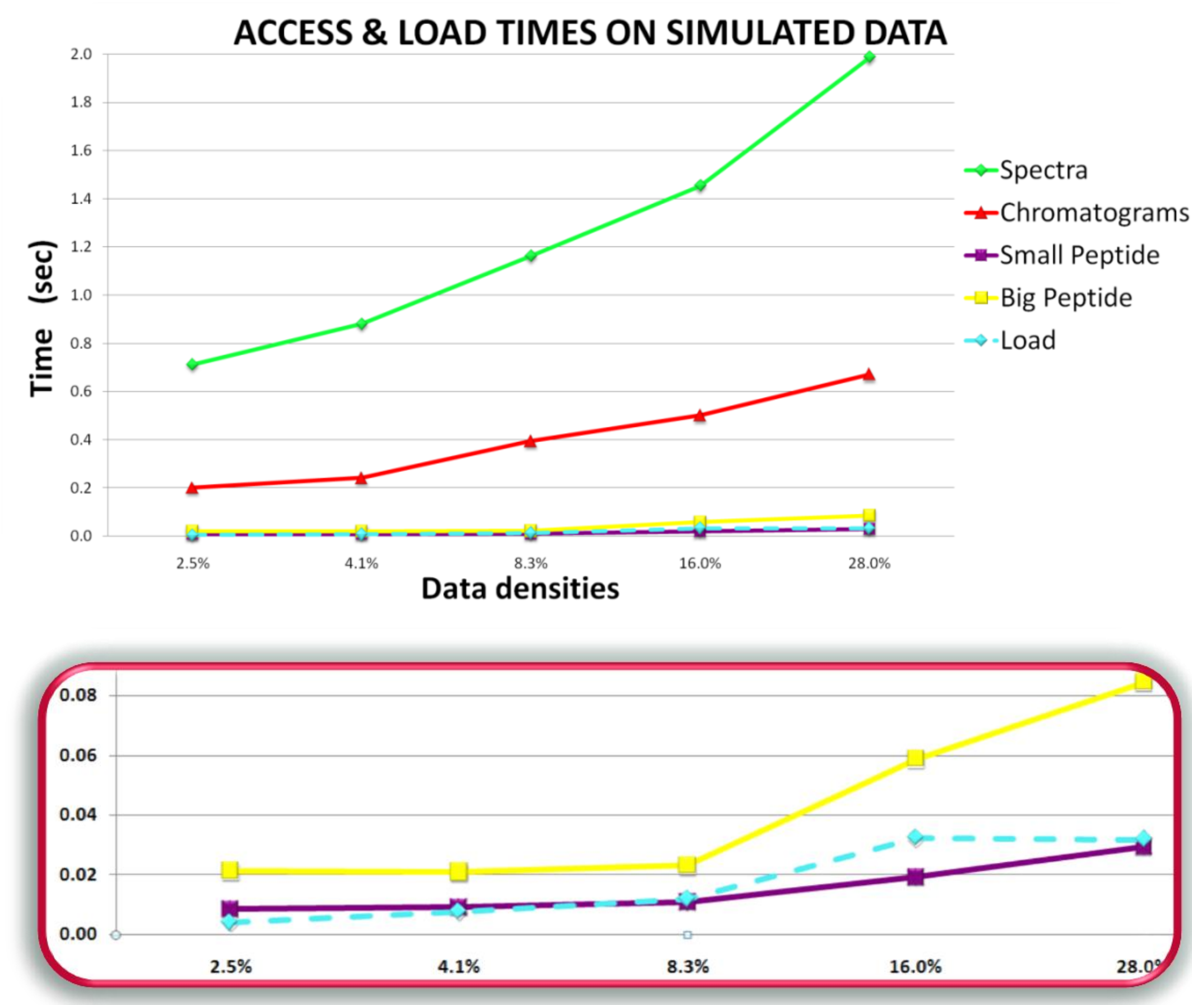


FIGURE 6-9 EVALUATION OF MZRTREE SCALABILITY ON INCREASING DATASET DENSITIES AS REGARDS THE LOAD TIME AND ACCESS TIMES ON DIFFERENT KIND OF RANGE QUERIES.AS CAN BE SEEN FROM THE ZOOMED IMAGE MZRTREE IS FAIRLY SCALABLE AS REGARDS ACCESS AND LOAD TIME: AS DATA DENSITY INCREASES BY A FACTOR OF 10, THE ACCESS TIME INCREASES ONLY BY A FACTOR OF 4 IN THE WORST CASE, WHILE THE LOAD TIME IS ALMOST CONSTANT.

6.5 DISCUSSION

In this chapter we described mzRTree, a scalable and memory efficient spatial structure for storing and accessing LC-MS data, which features efficient construction time and faster range query performance, compared to other known and widely used data structures.

Experimental results and the inherent scalability of the underlying R-tree structure suggest that mzRTree is suitable for high density/large size proteomics data, such as profile data, considered as the most informative and hence the most suitable to tackle quantification aims [23]. At present, profile data size reaches some GBs, but it is expected to further increase, as far as instrument accuracy and resolution increase: even a narrow range of m/z values can be challenging to manage when analyzing these data. Thus, the adoption of mzRTree for data storage could make profile data accessible for analysis purposes: it prevents out-of-memory errors, often occurring with huge profile proteomics datasets, and reduces the need for (and the costs of) extraordinary computational infrastructures and their management. Actually, profile data are often the only data source rich enough to perform a meaningful analysis, e.g., in quantitative proteomics based on stable isotope labeling . However, costs involved with profile data handling often outweigh their benefits. mzRTree could revert this relationship.

Several research questions remain open. The efficiency of mzRTree depends on several design choices, including the degree of the internal nodes and the way the bounding boxes are mapped to the leaves of the tree. The design space for mzRTree should be fully explored in order to identify the best choices. Moreover, when dealing with huge raw datasets mzRTree may not fit in RAM. In that case, the tree must reside on hard disk and the size of the internal nodes should be adapted to match the minimum block size used in disk-RAM data movements. Other solutions based on indexing structures alternative to the R-tree employed by mzRTree (e.g., those surveyed in [90], including the kd-tree used in [68]) should be considered and compared to mzRTree. Finally, it is interesting and potentially useful to investigate effective ways to further integrate all additional information needed for regulatory submission into mzRTree.

Recently, mzRTree was proposed to the PSI community as a valuable computational support to existing standards. At the moment a project is under development regarding the possibility of making use of mzRTree to realize a new open data format for efficient data handling in collaboration with foreign researchers involved in the development of PSI data formats and ontologies.

7 QUANTIFICATION: THE 3DSpectra SOFTWARE

In this chapter it will be presented 3DSpectra, an innovative quantification software for LC-MS labeled profile data developed under MATLAB (2008a, The MathWorks) environment. In order to achieve reliable peptide quantifications, the algorithm developed during this PhD research project exploits both the 3D LC-MS data resolving power and the informative content carried by profile data. In addition, it keeps down computational costs both for data handling and quantification. Indeed, in contrast to other available tools, 3DSpectra features optimized data handling by means of mzRTree [92], and a hybrid 2D and 3D data analysis approach. The 2D signal processing on chromatograms and spectrograms is coupled to a 3D peaks' borders recognition method. In this last step, 3DSpectra, by means of the Expectation-Maximization (EM) approach, fits the isotopic distribution shaped by a bivariate Gaussian Mixture Model (GMM) including a noise component on 3D peptide data. The estimated GMM is used to statistically define the boundaries of the peptide isotopic distribution. 3DSpectra substantially improves quantification efficiency compared to the state of the art software, and features the same good quantification accuracy and reliability. Furthermore, 3DSpectra achieves a significantly higher reproducibility of its peptide quantifications across experimental replicates. In addition, it showed high linearity and reliability. Here, we present 3DSpectra, a reliable and accurate quantification strategy, which provides significantly wide and reproducible proteome coverage.

Mass spectrometers can generate tremendous amounts of data, whereas accurate and reliable quantification is a rather computational intensive task. Thus, the analysis of the whole data would be a waste of computational resources. Consequently, 3DSpectra performs a local analysis focused on identified peptides where each peptide is analyzed separately by the software. In order to accomplish such a local analysis 3DSpectra creates, as a preliminary step, a metadata structured collection, called peptide library, containing information about the identified peptides (Figure 7-1). LC-MS data have first to be searched using the search engine of choice (e.g., Mascot, Sequest, X!Tandem [93], etc). Then, the a priori information has to be stored in a metadata file following a strictly defined schema, which will be provided to the user. This file is given as input to 3DSpectra. Afterwards, a peptide library, is automatically generated by 3DSpectra starting from the metadata file. It will be used during elaboration to retrieve peptide metadata, while the data are stored using mzRTree (see Chapter 6) to allow for an efficient data access during data analysis.

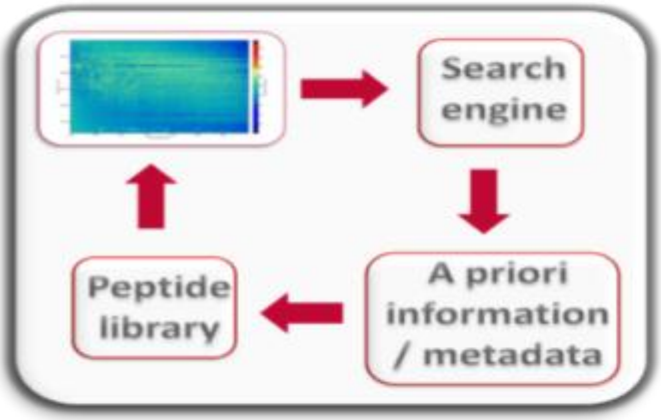


FIGURE 7-1 THE FIGURE ILLUSTRATES THE PEPTIDE LIBRARY. LC-MS DATA ARE SEARCHED BY THE PREFERRED SEARCH ENGINE. ITS RESULTS NEED TO BE STORED IN A FILE FOLLOWING A CERTAIN SCHEMA. STARTING FROM THIS FILE THE PEPTIDE LIBRARY IS BUILT UP. IT WILL BE USED DURING ANALYSIS TO RETRIEVE DATA.

Therefore, 3DSpectra exploits “a priori” information provided by search engines to analyze only areas of interest (i.e., data sub-matrices related to identified peptides), which are efficiently accessed using mzRTree. This is done iteratively by the algorithm, analyzing one identified peptide per iteration.

At every iteration of the algorithm the following steps take place, respectively:

1. **Metadata retrieval for local peptide analysis**. The peptide library is used to retrieve the metadata necessary for the subsequent analysis.

2. **Optimized data access via mzRTree**. The required data are loaded in memory using mzRTree.

3. **Main isotopic peak detection**. The algorithm detects the main isotopic peak of the peptide distribution.

4. **3D isotopic distribution model**. The theoretical isotopic distribution of the main peaks is modeled in the three dimensional space by a Gaussian Mixture Model (GMM) and fitted on data by the Expectation Maximization (EM) algorithm.

5. **Recognition of the isotopic distribution borders**. Peak borders are defined making use of the GMM.

6. **Processing and ratio computation.** Quantitative values for the peptides are calculated and the ratios of the differentially labeled peptides are computed.

In the following, 3DSpectra main steps will be exhaustively described, while a schematic representation is depicted by the flowchart in Figure 7-2.

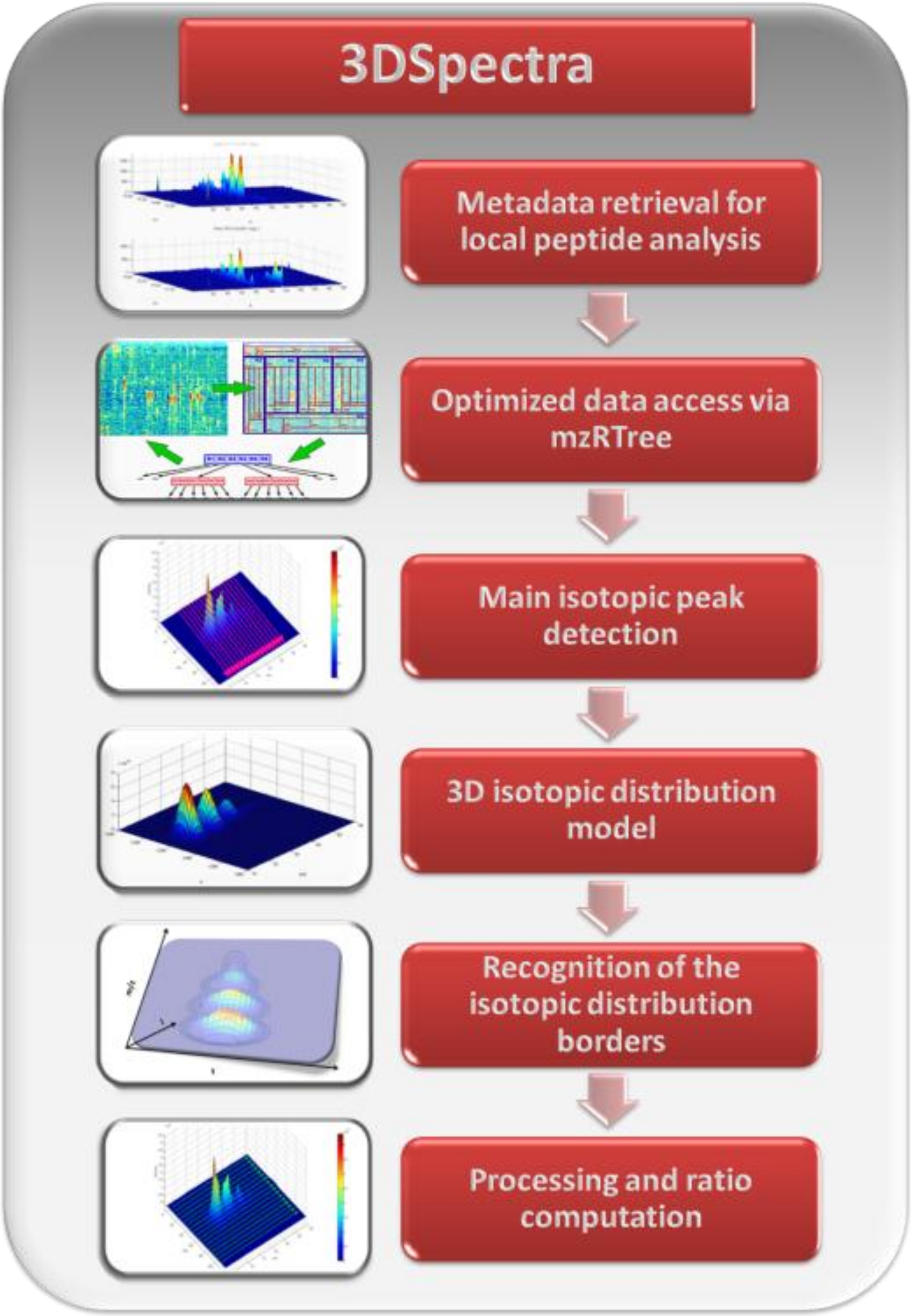

3DSpectra
Metadata retrieval for local peptide analysis
Optimized data access via mzRTree
Main isotopic peak detection
3D isotopic distribution model
Recognition of the isotopic distribution borders
Processing and ratio computation

FIGURE 7-2 THE ABOVE FIGURE ILLUSTRATES 3DSPECTRA WORKFLOW VISUALIZING THE MAIN STEPS OF 3DSPECTRA'S ALGORITHM AS REPORTED IN THE MAIN TEXT.

7.1.1 METADATA RETRIEVAL FOR LOCAL PEPTIDE ANALYSIS

The peptide library is used to retrieve peptide metadata and compute all information necessary to retrieve the data associated to the peptide under analysis and its isotopic partner.

In particular, the information about the theoretical distribution and peptide charge status is used to compute the m/z range for the sub-matrix of interest, coupled to the information on the identified elution time.

Then, in order to recognize the data sub-matrix associated to the isotopic partner of the peptide under analysis, the information regarding labeling is used to compute the position of the isotopic partner along the m/z dimension, which is shifted because of the label, whereas co-elution is hypothesized along the chromatographic dimension.

Figure 7-3 shows the data associated to two peptide sub-matrices, relative to an isotopic pair. Co-elution can be noticed, as well as the m/z shift due to the labeling.

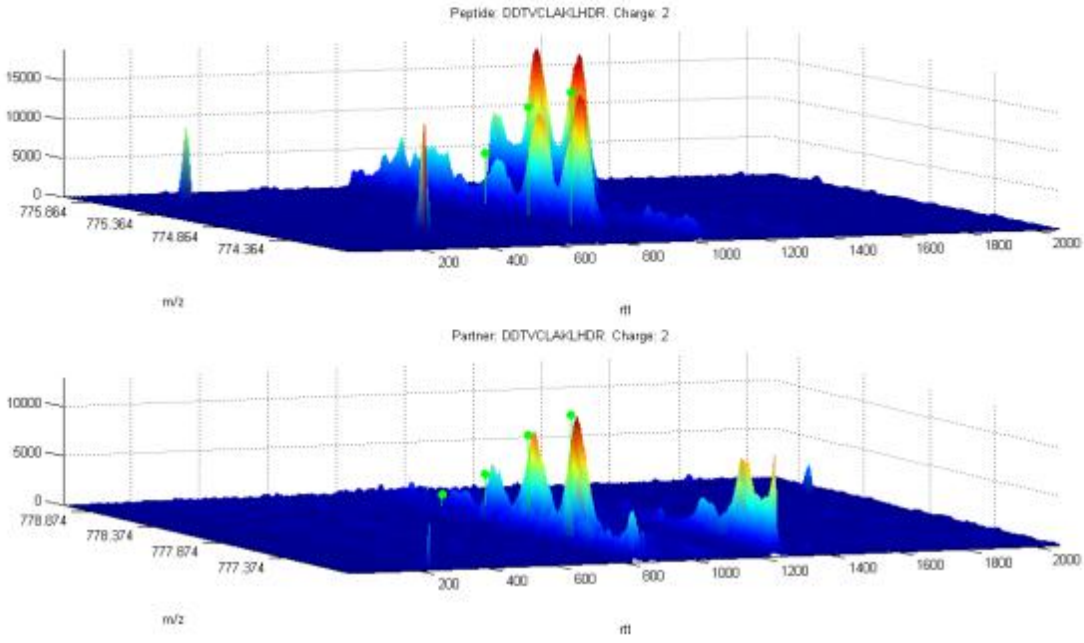


FIGURE 7-3 THE FIGURE VISUALIZES AN ISOTOPICALLY LABELED PAIR (PEPTIDE, PARTNER). GREEN DOTS ARE SHOWING WHERE 3DSPECTRA PREDICT THE ISOTOPIC PEAKS BELONGING TO THE DISTRIBUTION, BASED ON THE METADATA GATHERED FROM THE PEPTIDE LIBRARY.

7.1.2 OPTIMIZED DATA ACCESS VIA MZRTREE

The peptide library allows 3DSpectra to perform a local peptide analysis. However, repeated data accesses are computationally demanding since standard data formats, like mzXml/mzMl (*18*, *19*) (see Standard data formats), have been developed for data exchange, not for computation [63]. Thus, the required data associated to the peptide under analysis and its isotopic partner are accessed by means of mzRTree, which allows efficient data access even on huge data files. For more details see DATA HANDLING: THE MZRTREE DATA STRUCTURE.

This approach, embedded in 3DSpectra, ensures efficiency on data accesses for chromatograms, spectra and peptides; scalability to data density and size; hard disk space efficiency.

7.1.3 MAIN ISOTOPIC PEAK DETECTION

Once the peptide sub-matrix has been loaded, 3DSpectra combines a 2D and a 3D approach to process the retrieved data: a 2D signal processing on both chromatograms and spectrograms is coupled to a 3D peaks borders recognition method, based on a statistical model of the peptide isotopic distribution.

As a first step, a sum of Gaussians model is fitted to each chromatogram belonging to the sub-matrix of interest by Non Linear Least Squares (NLLS) (see Figure 7-4). The model used for fitting chromatographic peaks is:

$$y = \sum_{k=1}^{N} a_k e^{\left[-\left(\frac{x-m_k}{\sigma_k}\right)^2\right]} \qquad \text{7-1}$$

where a_k is the amplitude, m_k is the kth peak centroid and σ_k is the peak width of the kth Gaussian component. The maximum number of Gaussians N is 4 (for more details, see Implementation). Then, for each chromatogram, we select the Gaussian centroid associated to the maximum amplitude among the 4 Gaussians. The mode and the median of these centroid values extracted from all chromatograms belonging to the range of interest are computed. The new estimate is deemed to be reliable if the mode is within a 0.5 fold change from the median and it differs less than a user-definable threshold, defining a range of interest, compared to the elution time value retrieved from the peptide library. In that case, it substitutes the value provided by metadata as the true elution time.

Then, 3DSpectra recognizes the peptide distribution maximum peak looking for it on a 2D window defined by a narrow range along retention times centered on the newly estimated elution time and a m/z range equal to the peptide distribution width along the m/z dimension. The indexes (m/z*, t*) relative to this maximum are then used as a starting point for the fit of the 3D peptide distribution model used to define peptide peak borders in next step.

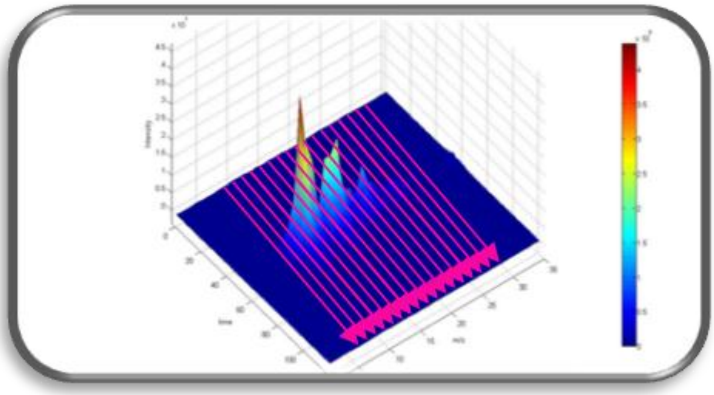


FIGURE 7-4 IN ORDER TO DETECT THE MAIN PEAK OF THE ISOTOPIC DISTRIBUTION 3DSPECTRA FITS A GAUSSIAN MODEL ALONG THE CHROMATOGRAPHIC DIMENSION. EACH CHROMATOGRAM IS FITTED BY ONE OR FOUR GAUSSIANS (PINK ARROWS) AND ITS MAXIMUM PEAK IS RECOGNIZED AS THE TALLEST ONE. THEIR MODES WILL BE USED TO EVALUATE THE ELUTION TIME.

7.1.4 3D ISOTOPIC DISTRIBUTION MODEL

In this step, a 3D isotopic distribution model is created by using a Finite Mixture Modeling (FMM) approach [94]. In fact, we fit a bivariate Gaussian Mixture Model (GMM) in (m/z – retention time) plane to the peptide isotopic distribution detected in the previous step. Initial parameters for the GMM are derived from the m/z and elution time computed in the main isotopic peak detection step (m/z*, t*) combined to “a priori” information on the theoretical isotopic distribution of the peptide under analysis. In order to exploit the “a posteriori” information carried by the data, a maximum likelihood (ML) estimation of the model parameters is performed by means of the expectation maximization (EM) algorithm. In the following we briefly show the main sub-steps involved in the ML estimation, and its solution via EM.

THE GAUSSIAN MIXTURE MODEL

The physical phenomenon generating LC-MS data is stochastic. Usually, LC-MS are referred to as signal intensities, but they actually are ion counts. In other words, LC-MS data correspond to the histogram of the real observations, i.e., the ions detected by the MS detector. Therefore, the entity ion could be seen as a random vector $x_i = \left(\frac{m}{z}_i, rt_i\right),\ i = 1, \dots, n$.

Thus, its probability density function (PDF) can be estimated from the LC-MS signal. Biochemistry teaches that such a distribution should follow some theoretically known shape factors ("a priori" knowledge), MS data gives additional information ("a posteriori" knowledge). Consequently, the PDF can be modeled on data using the FMM approach, where a maximum likelihood (ML) estimation of the PDF parameters of a GMM is performed by means of the EM algorithm (see Figure 7-5).

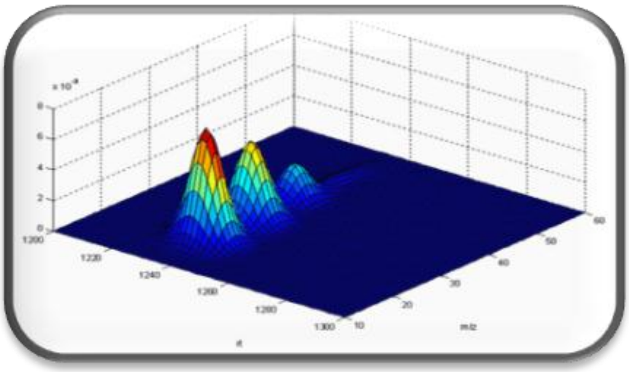


FIGURE 7-5 THE FIGURE SHOWS THE PDF ASSOCIATED TO A GMM THAT SHAPES THE ISOTOPIC DISTRIBUTION OF A PEPTIDE. THE GMM PDF CAN BE SEEN AS THE NORMALIZATION OF THE LC-MS SIGNAL, WHICH IS ITS HISTOGRAM.

We assume that data vectors $X = \left\{x_1,\ x_2, \dots, x_n \middle|\ x_i = \left(\frac{m}{z}_i, rt_i\right), \forall\ i = 1, \dots, n\right\}$ are independent and identically distributed with distribution p, whose parameters are represented by Θ. Thus, recalling the maximum likelihood estimation principle:

$$p(X|\Theta) = p(x_1, x_2, \dots, x_n|\Theta) = \prod_{i=1}^{n} p(x_i|\Theta) = L(\Theta|X) \tag{7-2}$$

where $p(X|\Theta)$ is equal to the likelihood function $L(\Theta|X)$ of the parameters Θ given the data X. $L(\Theta|X)$ is a function of the parameters Θ where the data X are fixed. The ML parameters estimate $\hat{\Theta}$ is given by the maximization of the likelihood function $L(\Theta|X)$:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\ L(\Theta|X) \tag{7-3}$$

Or, equivalently, by minimizing the $-log(L(\Theta|X))$, which is analytically and numerically more convenient:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} -log(L(\Theta|X)) \tag{7-4}$$

Depending on the form of $p(X|\Theta)$ the parameters estimation could be from easy to analytically intractable. Data on m/z dimension can be described by a sum of Gaussians distribution whose shape factors are defined by the theoretical isotopic distribution of the peptide [95]. Therefore the peptide distribution can be modeled as a probabilistic bivariate Gaussian Mixture Model:

$$p(X|\Theta) = \sum_{k=1}^{N} \alpha_k p_k(X|\theta_k) \tag{7-5}$$

where the parameters are $\Theta = (\alpha_k, \theta_k),\ k = 1, \dots, N$. Mixing proportions α_k are such that $\sum_{k=1}^{N} \alpha_k = 1$. Each p_k is a bivariate Gaussian PDF parameterized by $\theta_k = (\mu_k, \Sigma_k),\ k = 1, \dots, N$, where μ_k is the mean vector and Σ_k is the covariance matrix of the k^{th} Gaussian component. The GMM consists of as many Gaussian density components as is the number N of peaks considered for the theoretical isotopic distribution of the peptide. The $log(L(\Theta|X))$ to be maximized to estimate Θ is:

$$log\big(L(\Theta|X)\big) = log \prod_{i=1}^{n} p(x_i|\Theta) = \sum_{i=1}^{n} \log(\sum_{k=1}^{N} \alpha_k p_k(x_i|\theta_k)) \tag{7-6}$$

EXPECTATION MAXIMIZATION FOR THE GMM

The $log(L(\Theta|X))$ for the GMM is difficult to optimize because it contains the log of the sum. Here, the EM algorithm [96] was used, which is one of the most widely used in the computational pattern recognition community.

The hypothesis is that the observed data X is an incomplete set of data drawn from the distribution of which we want to estimate the parameters. The EM defines a complete dataset $Z=(X,Y)$ where $Y=\left\{y_i \in\{1,\ldots,N\} \mid x_i=\left(\frac{m}{z}_i, rt_i\right)\in G_k \leftrightarrow y_i=k, \forall\, i=1,\ldots,n\right\}$ is unknown and G_k is the k^{th} Gaussian component of the GMM. The PDF, which substitutes $p(X|\Theta)$, is then:

$$p(Z|\Theta)=p(X,Y|\Theta)=p(Y|X,\Theta)p(X|\Theta) \qquad 7\text{-}7$$

Therefore the log-likelihood function $\log(L(\Theta|X))$ is substituted with:

$$\log\big(L(\Theta|Z)\big)=\log\big(L(\Theta|X,Y)\big)=\log\big(p(X,Y|\Theta)\big) \qquad 7\text{-}8$$

The first step of EM algorithm, called Expectation step (E-step), estimates the expected value of the $\log\big(L(\Theta|Z)\big)$ with respect to the observed data X, the unknown data Y and the current parameter estimates $\hat{\Theta}=\Theta^{(i-1)}$. At the beginning, parameter estimates $\hat{\Theta}=\Theta^{0}$ can be extracted from some "a priori" information or simply random. The formulation for the expectation $Q\big(\Theta,\Theta^{(i-1)}\big)$ is:

$$\begin{aligned}Q\big(\Theta,\Theta^{(i-1)}\big) &= E\big[\log\big(L(\Theta|Z)\big)\big]=E\big[\log(p(X,Y|\Theta)|X,\Theta^{(i-1)}\big] \\ &=\sum_{\boldsymbol{y}\in Y} log\big(p(X,\boldsymbol{y}|\Theta)\big)p\big(\boldsymbol{y}\big|X,\Theta^{(i-1)}\big) \\ &=\sum_{\boldsymbol{y}\in Y}\sum_{i=1}^{n} log(\alpha_{y_i}p_{y_i}(x_i|\theta_{y_i}))\prod_{j=1}^{n}p(y_i|x_i,\Theta^{(i-1)})\end{aligned} \qquad 7\text{-}9$$

where Θ is a normal variable we are adjusting, X and $\Theta^{(i-1)}$ are known, Y is a random variable related to the unobserved data and its distribution is:

$$p\big(Y\big|X,\Theta^{(i-1)}\big) \qquad 7\text{-}10$$

which is the posterior probability of each GMM component with respect to each observation (i.e., ion). Notice that $\Theta^{(i-1)}$ are the parameters used to evaluate the expectation, whereas Θ are the parameters we are going to optimize in order to maximize the likelihood $L(\Theta|Z)$.

Indeed, the EM algorithm in a second step, called Maximization step (M-step), maximizes the expectation computed in the former step:

$$\Theta^i = \underset{\Theta}{\operatorname{argmax}}\ Q(\Theta, \Theta^{(i-1)}) \tag{7-11}$$

The E-step and M-step are iteratively repeated until a local maximum of the likelihood function is reached. For a GMM, the new estimates $\hat{\Theta}$ of the parameters based on the old estimates are as follows:

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} p(k|x_i, \Theta) \tag{7-12}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} x_i p(k|x_i, \Theta)}{\sum_{i=1}^{n} p(k|x_i, \Theta)} \tag{7-13}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} p(k|x_i, \Theta)(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{\sum_{i=1}^{n} p(k|x_i, \Theta)} \tag{7-14}$$

where k =1,…,N indicates the k^{th} Gaussian component and $p(k|x_i, \Theta)$ is the posterior probability of the k^{th} Gaussian component with respect to each ion. These update equations perform both E-step and M-step.

Since the local optimum of the likelihood function is strongly dependent on starting values, it is quite important to supply suitable EM starting parameters, which are Gaussians' centers (i.e., $\hat{\mu}_k$) and shapes (i.e., $\hat{\alpha}_k$ and $\hat{\Sigma}_k$). These values are extracted by the metadata stored in the peptide library, the theoretical isotopic distribution associated to the peptide under analysis and the main isotopic peak position estimated in the main isotopic peak detection step (see paragraph 7.1.3).

NUMBER OF COMPONENTS/ISOTOPES FOR THE GMM

One of the most important step for creating a good data model, both for clustering and GMM's parameters estimation, is to choose a suitable number of components: too few components fail to model the data accurately; too many components lead to an over-fit model with singular covariance matrices. If the number N of components of the GMM is unspecified, 3DSpectra determines an appropriate number of components, ranging from 2 to 5 Gaussians. This is achieved minimizing the Akaike information term (i.e., negative log-likelihood for the data with a penalty term for the number of estimated parameters). The Akaike Information Criterion (AIC) formulation is:

$$AIC = 2 \cdot NlogL + 2 \cdot m \qquad \text{7-15}$$

where m is the number of estimated parameters and $NlogL$ is the optimum negative log-likelihood for the estimated parameters.

7.1.5 RECOGNITION OF THE ISOTOPIC DISTRIBUTION BORDERS

Once the GMM's parameters have been estimated, the isotopic distribution borders can be defined in a statistical way in order to remove spurious ions. This borders recognition allows to determine which ions belong to the isotopic peptide distribution, hence should be quantified, and which do not. It consists of 2 sub-steps, which yield two different conditions to be both verified for the creation of a signal mask finally applied to the data matrix. These sub-steps are named:

1. 3D borders recognition
2. Noisy component identification

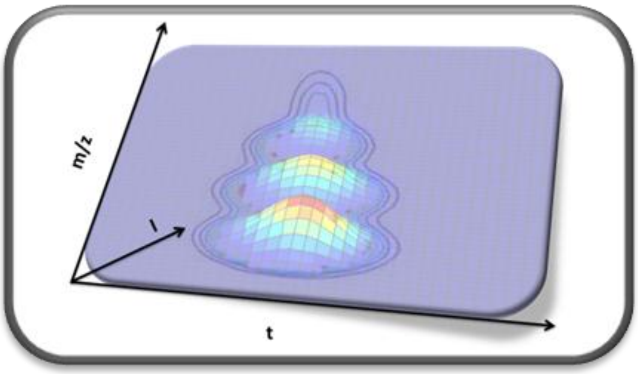


FIGURE 7-6 THE FIGURE SHOWS THE PDF ISO-DENSITY CURVES DEFINING THE BORDERS OF THE ISOTOPIC DISTRIBUTION. OUTLYING DATA ARE DISCARDED FROM SUBSEQUENT ANALYSIS. SYMBOL I REPRESENTS THE PDF VALUE, WHILE SYMBOL T REPRESENTS THE RETENTION TIME VALUE.

3D BORDERS RECOGNITION

In a first step, the borders of the isotopic distribution are identified by the GMM PDF iso-density curves (see Figure 7-6). The density value was empirically chosen (here, it is set to 0.0001) and ensures a conservative approach, that is to say keeping as many ions as possible. Then the first condition is that only data inlying the borders will be kept after filtering by signal mask.

NOISY COMPONENT IDENTIFICATION

In order to remove the noise from the ion counts, the GMM is used to recognize spurious ion counts deemed to be noise.

To accomplish this task data the GMM is used for clustering data. Indeed, in the literature, GMMs are often used for data clustering [97,98]: each Gaussian component of the fitted model corresponds to one cluster. Every observation (i.e., ion) in data is assigned to a cluster by choosing the component of the GMM with the largest posterior probability (see paragraph 3D isotopic distribution model, equation 7-10).

Then, 3DSpectra identifies a cluster (i.e., one of the Gaussian components) of spurious data or noise among all clusters associated to the GMM. The main features of the noise component are:

1. to cluster many ions,
2. to have few counts per ion,
3. to feature a big variance,
4. to be unaligned along the chromatographic dimension to the other Gaussians belonging to the GMM, in contrast to the other co-eluting components, which are clustering the peptide ion counts.

3DSpectra identifies the Gaussian component that satisfies the greatest number of these properties and recognizes it as the spurious one. The posterior probability of belonging to such a cluster is used to discard from any subsequent analysis noisy counts, that is to say ion counts having a posterior probability of belonging to the noisy Gaussian component higher than 0.9 (see Figure 7-7 panel (f)). This is the second condition for the signal mask.

Finally, the Boolean signal mask is defined merging the two above conditions and it is applied to the data matrix prior to further analysis (see Figure 7-7 panel (e)). Outlying data or data belonging to the noise component are discarded from subsequent analysis.

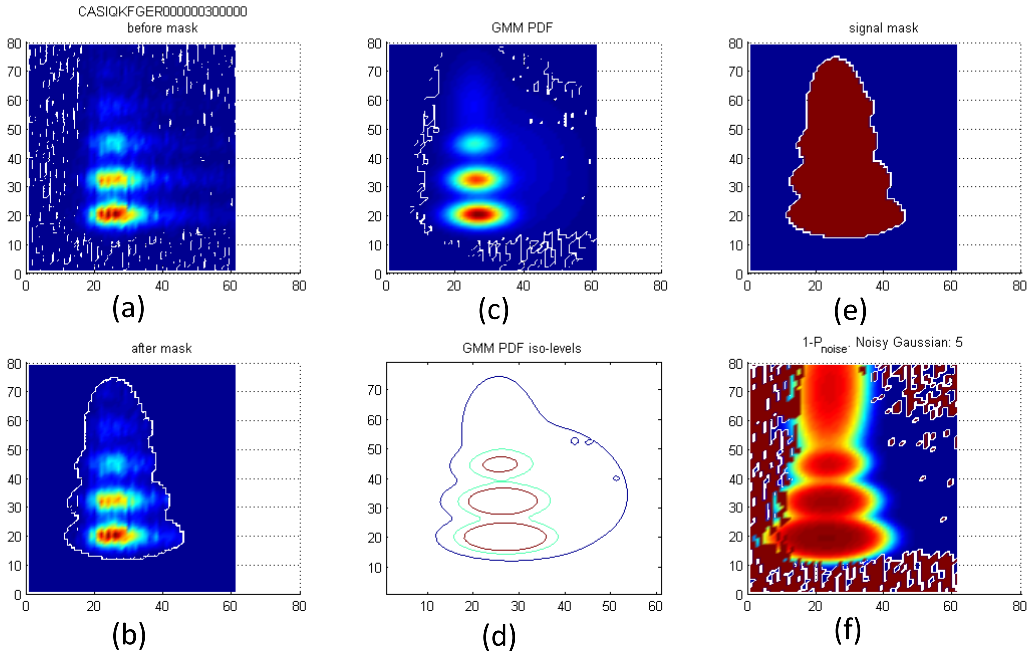


FIGURE 7-7 THE FIGURE ILLUSTRATES THE PEAK'S BORDERS RECOGNITION STEP EMBEDDING THE REMOVAL OF THE DATA BELONGING TO THE GMM COMPONENT ASSOCIATED TO THE NOISE (HERE, THE 5TH). IN PANEL (A) THE ORIGINAL SIGNAL IS REPRESENTED. IN PANEL (B) THE MASK HAS BEEN APPLIED. THE GMM PDF IS PLOTTED IN PANEL (C) AND IT CAN BE NOTICED THAT THE GMM CAN FOLLOW THE ELUTION PROFILE TO A GREAT EXTENT. IT IS CLEAR ALSO IN PANEL (D), WHERE THE PDF ISO-CURVES ARE PLOTTED. IN PANEL (E) THE SIGNAL MASK IS SHOWN, WHILE PANEL (F) ILLUSTRATES THE PROBABILITY OF NOT BELONGING TO THE NOISY COMPONENT (THE DARK RED SIGNAL IS DUE TO NO DATA).

7.1.6 PROCESSING AND RATIO COMPUTATION

After the 3D peaks' borders recognition step has been accomplished and the mask defining peaks' borders has been applied, only the data belonging to the peptide isotopic distribution are left. Spectra are smoothed using the Savitzky and Golay least-squares digital polynomial filter [38] along the m/z dimension. Then, on every spectrum, after grouping together the distribution peaks and summing the intensities belonging to each isotopic peak, a 2D peptide isotopic distribution model is fitted via Weighted Linear Least Squares (WLLS). The information gathered from metadata is used to weight each isotopic peak contribution to the abundance estimate. The weights are given by the probability of every isotopic peak in the theoretical model. Further on, for the sake of simplicity, the abundance estimate will be referred to as the volume under the curve (VUC) of the peptide distribution. From the WLLS fit we obtain a matrix, made of a number of chromatograms N, which is equal to the number of isotopic peaks considered for the distribution model. After that, N total ion currents values are extracted summing all intensities under the N chromatograms separately. In such a way, a spectrum made up of N values is achieved. On this final spectrum the 2D theoretical distribution is fitted once more via WLLS. The weights are given by the probability of every isotopic peak in the theoretical model. Then, the same weights are used for the quantification of VUC as weighted sum of the isotopic peaks contributions. The relative quantification is computed as the ratio of the peptide's VUC to its isotopic partner's one. An empirical reliability score, or weight, associated to each ratio is provided by the correlation between the data of the peptide and those of its labeled partner. Finally, to obtain results and statistics on the computed ratios, outlier removal can be performed by either Grubbs test or a MASPECTRAS built-in method.

7.2 IMPLEMENTATION

3DSpectra's algorithm is implemented in MATLAB and it is available upon request (3DSpectra@dei.unipd.it). In this paragraph, some details about its implementation are given.

7.2.1 METADATA RETRIEVAL FOR LOCAL PEPTIDE ANALYSIS

The peptide library is automatically generated by 3DSpectra starting from the metadata file path by means of the *library*(filePath) function. It works properly only if the metadata file follows a strictly defined schema, which is provided with the software itself. The peptide library variable is saved in a *.mat* file, which is loaded at the beginning of 3DSpectra execution.

7.2.2 OPTIMIZED DATA ACCESS VIA MZRTREE

In order to allow efficient and flexible data accesses, 3DSpectra is provided together with a data access toolbox enabling to retrieve data by range queries.

By default, data access is performed by the mzRTree default range query method. mzRTree can be automatically created by the 3DSpectra built-in function *mzRTreeCreation*(file_mzXML, file_mzRTree), starting from the mzML (or, mzXML) file path. The mzRTree data structure is then stored in the file_mzRTree path.

Alternatively, the user can choose to use the mzML/mzXML standard format, accessing it by means of the Java Random Access Library (JRAP) [99].

7.2.3 MAIN ISOTOPIC PEAK DETECTION

The fit of the sum of Gaussians model on every elution profile along the temporal dimension is implemented by means of the *fit*(retTimes,ionCounts,libname) function from the Curve Fitting Toolbox. It fits the data in the column vectors retTimes and ionCounts using the library model specified by libname, which is set to gauss4 (i.e., sum of 4 Gaussians model). Default settings are used.

7.2.4 3D ISOTOPIC DISTRIBUTION MODEL

The isotopic distribution shaped by the GMM is fitted to peptide data using the *gmdistribution.fit*(X,k) function from the Statistics Toolbox, which implements the Expectation Maximization (EM) algorithm. It outputs an object of the *gmdistribution* class containing maximum likelihood estimates of the parameters of the Gaussian mixture model with k components for data in the n-by-d matrix X, where n is the number of observations and d is the dimension of the data.

In particular, the gmdistribution.fit method assumes a collection of samples from the mixture are observed rather than an aggregate representation of the samples, such as the histogram. Since the observed mixture is the LC-MS signal, it gives an aggregate representation of samples. Thus, we need to compute the collection of samples that generated it. Such operation is computationally very demanding under MATLAB and in order to optimize it, a C++ source file has been compiled and linked into a shared library called a binary MATLAB Executable (MEX) file.

The theoretical isotopic distribution parameters are computed making use of some MASPECTRAS built-in methods which have been embedded in the implementation in a Java executable library.

7.2.5 RECOGNITION OF THE ISOTOPIC DISTRIBUTION BORDERS

To recognize the GMM PDF iso-density curves we used the *pdf*(gmm,X) function of the *gmdistribution* class. It returns a vector y of length n containing the values of the PDF for the *gmdistribution* object gmm, evaluated at the n-by-d data matrix X, where n is the number of observations and d is the dimension of the data.

Data clustering was implemented by the *cluster*(gmm, X) function from the *gmdistribution* class: the method assigns a cluster to each observation in the n-by-d data matrix X, where n is the number of observations and d is the dimension of the data, into k clusters determined by the k components of the Gaussian mixture distribution defined by gmm. It returns a n-by-1 vector of indexes, idx, where idx(I) is the cluster index of observation I referring to the component of the GMM with the largest posterior probability, weighted by the component probability.

The probability of each ion count of belonging to the noise component is estimated employing the *posterior*(gmm,X) function from the *gmdistribution*. It returns P, the posterior probabilities of each of the k components in the Gaussian mixture distribution defined by gmm for each observation in the data matrix X. P is a n-by-k matrix, with P(I,J) the probability of component J given observation I. X has n-by-d size, where n is the number of observations and d is the dimension of the data.

7.2.6 PROCESSING AND RATIO COMPUTATION

To implement the smoothing of spectra and chromatograms using the Savitzky and Golay method, we used the *mssgolay*(x, ionCounts) MATLAB function from the Bioinformatics Toolbox. It smoothes raw noisy signal data featuring peaks using least-squares polynomial. The x vector consists of separation-unit values. The ionCounts parameter is a vector of intensity values.

The theoretical isotopic distribution model is fitted on data by means of Weighted Linear Least Squares (WLLS), implemented in the *lscov*(A,b,w) MATLAB function. It computes a weighted least-squares (WLS) fit when provided with a vector of relative observation weights, w. It returns x, the weighted least squares solution to the linear system A*x = b, that is, x minimizes (b - A*x)'*diag(w)*(b - A*x) and here is a scalar. Matrix A is a vector made of the theoretical relative intensities in the isotopic distribution. The weights w are the probabilities of each isotopic peak.

The correlation reliability score, or weight, associated to each ratio is computed by the *corr2*(A, B). It computes the 2-D correlation coefficient between A and B, where A and B are the data matrices of the same size associated respectively to the peptide and its labeled partner.

Outlier removal is performed by either Grubbs test or a MASPECTRAS built-in method which have been embedded in the proposed implementation. The MASPECTRAS method was linked into a Java library.

In order to allow visual inspection a function for the automatic visualization of every pair (peptide, partner) has also been implemented.

Results are stored both in a MATLAB workspace variable and in an Excel file; regression lines of light to heavy volumes are printed to a postscript file.

3DSpectra can be compared to any other software, whose results are stored in an Excel file compliant to a well-defined schema; the compared regression lines will also be printed to a postscript file automatically.

Moreover MATLAB allows to browse the results variable through its visual editor, where it is possible to see how the variable is structured and which are the stored values. The variable has a field for every relevant information related to the analyzed peptide: the peptide sequence, its charge, its index to retrieve additional metadata from the peptide library (e.g., its labeling status, elution time, etc), the estimated quantification ratio, the VUC of both the peptide and its partner, the experimental replicate where the peptide has been found, the correlation value cited above. If the quantification ratio has been computed starting from multiple peptide occurrences with different charges, all of them are reported in the charge field, and the corresponding library indexes appear in the index field.

7.3 PERFORMANCE ASSESSMENT

In this section it is described how 3DSpectra performance were evaluated using a controlled dataset from a preceding study [100]. As previously described (see Chapter 5), it consists of real profile data from a controlled mixture of ICPL-labeled proteins (bovine serum albumin (UniprotKB: P02769), human apotransferrin (UniprotKB: P02787) and rabbit phosphorylase b (UniprotKB: P00489)). They were mixed at seven different light to heavy ratios (1:1, 1:2, 1:5, 1:10, 2:1, 5:1, 10:1) in triplicates. Acquisition was run in enhanced profile mode for survey scans to gain higher mass accuracy using a Finnigan LTQ linear ITMS (Thermo Electron) equipped with HPLC-NSI source. Published quantification results show that ASAPRatio (MASPECTRAS implementation) reaches the best performance compared to MSQuant and PepQuan (Bioworks 3.2, Thermo Electron). Therefore, we compared 3DSpectra to ASAPRatio only. In order to obtain comparable quality parameters, both methods used the same set of peptide identifications as starting point.

The quality parameters, which have been chosen for assessing quantification performance, are:

1. Accuracy, i.e., the ability to quantify peptide ratios with an accurate estimate.
2. Precision, i.e., the ability to quantify peptide ratios with both a small standard deviation and a small coefficient of variation.
3. Efficiency , i.e., the number of quantified peptides.
4. Reproducibility, i.e., the ability to quantify the same peptide across experimental replicates.
5. Reliability, i.e., the ability to reliably quantify peptide ratios featuring linearity across the dynamic range[2].

[2] i.e., the range of variation of the light and heavy VUC quantities used for the computation of the ratios themselves.

Quality parameters were evaluated as follows:

1. For assessing quantification accuracy, we estimated the mean of all quantification ratios for each dataset.
2. To evaluate quantification precision, we computed the standard deviation and coefficient of variation (i.e., the percentage ratio of the standard deviation to the mean) of all quantification ratios for each dataset.
3. To validate quantification efficiency, we compared the total number of peptide ratios provided by both methods after outlier removal.
4. For assessing quantification reproducibility, we analyzed quantification ratios provided by both methods across the three experimental replicates on a set of commonly quantified peptide sequences. This set is given by the intersection of all peptide sequences quantified by the two methods. Every peptide sequence could be associated to at most three quantification ratios, each associated to one peptide occurrence per replicate. The ideally reproducible algorithm would quantify every peptide sequence three times: one per replicate.
5. To validate the quantification reliability we performed the analysis of Deming regression lines [101-103] between light and heavy abundances. Deming regression was chosen since it accounts for errors both on x and y observations. In order to make them comparable, the regression lines were evaluated:

   5.1. only on the common peptides quantified on the same replicate by both methods,

   5.2. and all peptide abundances were normalized to the maximum value for each algorithm.

7.4 RESULTS

Results for the above mentioned quality parameters evaluated on three different subsets of quantified peptides considered during this comparative analysis are reported in Table 7-1, Table 7-3, and Table 7-4. The last three rows of each table regards quantification accuracy and precision: mean, standard deviation (*SD*) and coefficient of variation (*CV*) of the ratios computed by both methods across all datasets are shown. All tables clearly demonstrate that 3DSpectra and ASAPRatio reach the same quantification accuracy and precision over all datasets and on all subsets of quantified peptides considered during this comparative analysis.

Results regarding quantification efficiency are reported in Table 7-1. It report the number of all quantified peptide occurrences across all experimental replicates (*Quantified peptides*) and the corresponding unique peptide sequences (*Unique pep seqs*) in the first and second row, respectively. Third row is the percentage of ASAPRatio to 3DSpectra Unique pep seqs values (*2D Coverage*). The first row of Table 7-1 demonstrates that 3DSpectra can quantify 2 to 4 times more differentially expressed peptide ratios, which are of key interest from a biological point of view, e.g., in biomarkers discovery. Moreover, compared to ASAPRatio, all quantification ratios by 3DSpectra yield to a much higher number of quantified unique peptide sequences (see Table 7-1, second and third rows). ASAPRatio can quantify indeed only around 22% to 48% of 3DSpectra unique peptide sequences quantifications for differentially expressed ratios, as reported by the third row of Table 7-1.

Efficiency	**1l:2h**		**2l:1h**		**1l:5h**		**5l:1h**		**1l:10h**		**10l:1h**		**1l:1h**	
	3D	**2D**	**3D**	**2D**	**3D**	**2D**	**3D**	**2D**	**3D**	**2D**	**3D**	**2D**	**3D**	**2D**
Quantified peptides	**98**	**39**	**118**	**38**	**99**	**38**	**108**	**34**	**122**	**23**	**116**	**23**	**111**	**85**
Unique pep seqs	**52**	**25**	**63**	**25**	**52**	**24**	**53**	**18**	**61**	**17**	**58**	**13**	**61**	**37**
2D Coverage	**48%**		**40%**		**46%**		**34%**		**28%**		**22%**		**61%**	
Mean ratio	0.53	0.54	1.95	1.99	0.26	0.27	4.94	4.23	0.16	0.13	8.55	9.08	1.14	1.06
SD	0.16	0.14	0.56	0.66	0.08	0.09	1.44	1.18	0.06	0.05	2.90	3.37	0.26	0.27
CV	29%	26%	29%	33%	29%	34%	29%	28%	42%	37%	34%	37%	23%	26%

TABLE 7-1 3DSPECTRA AND ASAPRATIO COLUMNS ARE RESPECTIVELY 3D AND 2D LABELED. “QUANTIFIED PEPTIDES” IS THE NUMBER OF ALL QUANTIFIED PEPTIDE OCCURRENCES ACROSS ALL EXPERIMENTAL REPLICATES. “UNIQUE PEP SEQS” IS THE NUMBER OF THE CORRESPONDING UNIQUE PEPTIDE SEQUENCES. “2D COVERAGE” IS THE PERCENTAGE OF ASAPRATIO TO 3DSPECTRA “UNIQUE PEP SEQS” VALUES. IT ALSO REPORTS MEAN, STANDARD DEVIATION (SD) AND COEFFICIENT OF VARIATION (CV) OF THE RATIOS COMPUTED BY BOTH METHODS ACROSS ALL DATASETS.

In particular, the new algorithm is advantageous for differentially expressed ratios, which are the most difficult to quantify, as far as the differential expression increases. In fact, ASAPRatio efficiency worsens at highly differentially expressed ratios, whereas 3DSpectra feature the same efficiency across all ratios. In conclusion, 3DSpectra achieves a significantly higher proteome coverage at the level of peptide quantification compared to ASAPRatio, especially for differentially expressed ratios.

To understand if the quantifications provided by 3DSpectra include those provided by ASAPRatio, we evaluated the overlap between the unique peptide sequences quantified by both methods (see Table 7-2). For differentially expressed ratios, 3DSpectra quantified on average 94% of all unique peptide sequences quantified by ASAPRatio, whereas ASAPRatio just 34% of those reported by 3DSpectra. 3DSpectra can quantify almost all unique peptide sequences quantified by ASAPRatio while ASAPRatio is able to quantify only one third of 3DSpectra's. Thus, 3DSpectra attains a much higher sequence coverage, which could be crucial for biomarkers discovery studies, as well as reproducibility.

Overlap	1l:2h		2l:1h		1l:5h		5l:1h		1l:10h		10l:1h		1l:1h	
	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D
Common pep seqs	24		21		22		17		17		13		31	
Unique pep seqs	52	25	63	25	52	24	53	18	61	17	58	13	61	37
Overlap	96%	46%	84%	33%	92%	42%	94%	32%	100%	28%	100%	22%	84%	51%

TABLE 7-2 3DSPECTRA AND ASAPRATIO COLUMNS ARE RESPECTIVELY 3D AND 2D LABELED. IT REPORTS THE NUMBER OF COMMONLY QUANTIFIED PEPTIDE SEQUENCES (COMMON PEP SEQS), THE TOTAL NUMBER OF UNIQUE PEPTIDE SEQUENCES QUANTIFIED BY EACH ALGORITHM (UNIQUE PEP SEQS) AND THEIR PERCENTAGE OVERLAP WITH THE NUMBER OF COMMONLY QUANTIFIED PEPTIDE SEQUENCES (OVERLAP).

Table 7-3 reports statistics related to the assessment of quantification reproducibility. The table shows the number of commonly quantified peptide sequences (*Common pep seqs*) as in the first row of Table 6-1 and the corresponding maximum number of peptide occurrences that can be found across the three experimental replicates (*Max # occurrences = 3 x Common pep seqs*) in the first and second row, respectively. Third row reports the actual number of quantified peptide occurrences among all possible occurrences across the three replicates (*Quantified peptides*). The fourth row shows the percentage coverage across the three replicates given by the ratio of *Quantified peptides* to *Max # occurrences* (*Replicate Coverage*). The *Replicate Coverage* parameter summarizes the information about the coverage offered by the two methods and sheds light on the much higher coverage of 3DSpectra across the replicates. We found out that 3DSpectra can quantify on average 84% of all possible peptide occurrences for differentially expressed ratios, whereas ASAPRatio only 54%. Thus, 3DSpectra achieves a significantly higher reproducibility of its peptide quantifications across experimental replicates, quantifying 30% more peptide occurrences than ASAPRatio.

Reproducibility	1l:2h		2l:1h		1l:5h		5l:1h		1l:10h		10l:1h		1l:1h	
	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D
Common pep seqs	24		21		22		17		17		13		31	
Max # occurrences	72		63		66		51		51		39		93	
Quantified peptides	58	38	47	31	48	36	49	32	43	23	33	23	73	69
Replicate Coverage	**81%**	**53%**	**75%**	**49%**	**73%**	**55%**	**96%**	**63%**	**84%**	**45%**	**85%**	**59%**	**78%**	**74%**
Mean ratio	0.55	0.54	2.08	2.06	0.27	0.27	4.56	4.34	0.15	0.13	9.49	9.08	1.11	1.11
SD	0.15	0.14	0.56	0.62	0.07	0.09	1.42	1.13	0.05	0.05	2.78	3.37	0.25	0.26
CV	28%	26%	27%	30%	27%	33%	31%	26%	36%	37%	29%	37%	23%	23%

TABLE 7-3 3DSPECTRA AND ASAPRATIO COLUMNS ARE RESPECTIVELY 3D AND 2D LABELED. THE TABLE ILLUSTRATES THE NUMBER OF COMMONLY QUANTIFIED PEPTIDE SEQUENCES (COMMON PEP SEQS), THE MAXIMUM NUMBER OF PEPTIDE OCCURRENCES ASSOCIATED TO COMMON PEP SEQS THAT CAN BE FOUND ACROSS THE THREE EXPERIMENTAL REPLICATES (MAX # OCCURRENCES = 3 X COMMON PEP SEQS), THE ACTUAL NUMBER OF QUANTIFIED PEPTIDES (QUANTIFIED PEPTIDES) AND THE COVERAGE GIVEN BY THE PERCENTAGE RATIO OF QUANTIFIED PEPTIDES TO MAX # OCCURRENCES (REPLICATE COVERAGE).

Results regarding the assessment of reliability by means of regression analysis are shown in Table 7-4. It reports the main parameters related to the linear model describing the light to heavy estimates relationship: the squared Pearson's correlation coefficients (R^2), the Root Mean Squared Error (*RMSE*). In order to make them comparable, the regression lines have been evaluated only on common peptides quantified on the same replicate by both methods (*Common peptides*, first row), which are associated to the reported number of uniquely commonly quantified peptide sequences (*Common pep seqs*, second row).

As a preliminary step to regression analysis we needed to verify linearity between light and heavy abundances for all datasets. Thus, we computed Pearson's correlation coefficients, which resulted to be 1% statistically significant. Both methods feature a strong linear relationship between light and heavy abundances (see Table 7-4, third row). There is no statistically significant difference among the methods, except that for the 10:1 ratio, where 3DSpectra shows a 5% significantly higher Pearson correlation coefficient. After that, we performed Deming regression and computed the RMSE to evaluate the quantification reliability (see Table 7-4, fourth row). The RMSE associated to 3DSpectra is on average smaller than the RMSE related to ASAPRatio, but overall they can be considered comparable.

Reliability	1l:2h		2l:1h		1l:5h		5l:1h		1l:10h		10l:1h		1l:1h	
	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D
Common peptides	**36**		**26**		**31**		**31**		**21**		**20**		**59**	
Common pep seqs	24		19		20		17		17		12		29	
R^2	**0.96**	**0.91**	**0.87**	**0.91**	**0.94**	**0.86**	**0.96**	**0.91**	**0.77**	**0.89**	**0.98**	**0.88**	**0.95**	**0.92**
RMSE	**0.06**	**0.09**	**0.09**	**0.08**	**0.07**	**0.11**	**0.05**	**0.06**	**0.08**	**0.08**	**0.04**	**0.06**	**0.07**	**0.08**
Mean ratio	0.56	0.55	2.06	2.15	0.27	0.28	4.70	4.38	0.14	0.14	9.66	9.51	1.11	1.11
SD	0.15	0.14	0.62	0.57	0.08	0.09	1.45	1.13	0.06	0.05	2.69	3.35	0.26	0.26
CV	27%	26%	30%	27%	28%	31%	31%	26%	41%	36%	28%	35%	23%	23%

TABLE 7-4 3DSPECTRA AND ASAPRATIO COLUMNS ARE RESPECTIVELY 3D AND 2D LABELED. PARAMETERS ARE REPORTED RELATED TO THE LINEAR MODEL DESCRIBING THE LIGHT TO HEAVY ESTIMATES RELATIONSHIP: THE SQUARED PEARSON'S CORRELATION COEFFICIENTS (R^2), THE ROOT MEAN SQUARED ERROR (RMSE). IN ADDITION, ALL STATISTICS ARE SHOWN, SUCH AS MEAN, SD AND CV.

Then, we assessed the correctness of the linear model by the Fisher-Snedecor (F)-statistic. The results were zero for both methods across all datasets, thus the linear model is an adequate solution to describe the relationship between light and heavy abundances, as previously predicted by the Pearson's correlation coefficients. In conclusion, the two methods feature the same quantification reliability.

The experimental results clearly demonstrated that 3DSpectra achieves significantly higher protein sequence coverage and reproducibility than ASAPRatio, and features the same quantification accuracy, precision and reliability.

7.5 DISCUSSION

3DSpectra is an innovative analysis algorithm for the quantification of LC-MS labeled data. It features an optimized data handling and an innovative peaks' borders recognition method, leading to outstanding results in terms of quantification efficiency and reproducibility, providing the same accuracy and reliability as the well-known ASAPRatio algorithm.

Quantification efficiency is critical for proteomics research since it plays a crucial role in biomarkers discovery studies: the wider the proteome coverage at the level of peptide/protein quantification, the higher the probability of discovering differentially expressed peptides/proteins among different biological conditions. In biomarkers discovery studies, quantification efficiency is as important as quantification accuracy. Indeed, differential expression to the reference sample is often considered meaningful if it is at least doubled. Furthermore, the more peptides related to a certain protein are quantified, the more reliable is the protein quantification.

Likewise, the quantification reproducibility could be pivotal as well. For instance, it could help classification algorithms in distinguishing differentially expressed peptides between control versus unhealthy samples, especially when several samples are available per every class.

The goal should be to increase the amount of reliably and reproducibly quantified peptides to raise the quality level of expression studies, and accordingly the confidence in correlated biological findings. Therefore, quantitative proteomics must focus on quantification efficiency, still ensuring a good accuracy and reliability, and as far as possible reproducibility.

Here, we evaluated 3DSpectra performance employing real profile data from a controlled mixture of Isotope Coded Protein Labels (ICPL)-labeled proteins mixed at different ratios in triplicates and acquired in enhanced profile mode. We showed that 3DSpectra quantifies, on differentially expressed ratios, 2 to 4 times more peptide ratios than ASAPRatio, resulting in a substantial improvement (100% to 300%) in quantification efficiency. Furthermore, the wider proteome coverage here comes with no tradeoff: 3DSpectra reached the same performance as ASAPRatio regarding quantification accuracy, precision and reliability, indeed. Moreover, 3DSpectra achieves a 30% higher reproducibility of its peptide quantifications across experimental replicates.

The obtained excellent results are deemed to be the effect of 2 main causes: 1) a 3D approach that minimizes the peptides overlapping; 2) a peak border recognition method that recognizes all ion counts belonging to the peptide isotopic distribution and estimates their probability of being noise. 3DSpectra is therefore able to reduce the number of misquantified peptides. In fact, 3DSpectra could quantify peptide hits, which would be eliminated in the outlier removal step of ASAPRatio as implemented in MASPECTRAS. Consequently, the amount of peptide ratios is substantially increased.

Here, 3DSpectra's quantification performance has been evaluated on low resolution data, where a major degree of uncertainty is associated to identification results because of the low mass accuracy. Therefore, the reported results can be considered as a worst case evaluation of the 3DSpectra algorithm. 3DSpectra's performance is expected to be enhanced by high mass accuracy datasets, and will be demonstrated in future work. Nonetheless, this dataset highlighted 3DSpectra's ability to efficiently and reproducibly quantify even low resolution data.

A common problem in the analysis of MS-based proteomics data is that only the more abundant peptides are usually covered by identification and, hence, quantification. Data related to the less abundant peptides are unlikely to be analyzed, eventually wasting their biological meaning. Thus, also the estimation of the quantification efficiency could be biased from the higher abundant proteins in the sample. This sample, being a simple controlled mixture of proteins, ensured that also less abundant peptide hits could be identified and thus quantified.

It would be interesting to evaluate the quantification efficiency making use of Selected Reaction Monitoring (SRM) data, where the experimental design is such that the acquired sample proteome is already known. Therefore, both the less abundant peptides and/or the peptides missed by search engines and those actually present in the sample but identified with a low confidence will be analyzed by the quantification software, since the identifications are "a priori" known. To our knowledge, an SRM dataset suitable for the assessment of quantification performance is still not available to the community in public repositories.

Future developments will focus on checking the conformity of the model to a broader range of proteomics MS-based data (e.g., SRM data), considering also possible suitable modifications to the model itself. Besides, the GMM approach would allow the association of a statistical reliability score or weight to each ratio based on the error estimates or confidence intervals on the parameters of the GMM. This step is extremely computationally demanding since the bootstrap approach is needed to estimate the standard error (error estimates for the GMM). That makes error estimates unfeasible at the moment. The optimization of this step and the whole 3DSpectra software could be an interesting additional improvement.

Further work is then needed to facilitate the import of identification results and the export of quantification results, adding support for the Proteomics Standards Initiative exchange data formats, i.e., mzIdentML [104] and, as soon as its final documentation will be released, mzQuantML [105].

CONCLUSIONS

In this thesis were described the state of art, the design and development of methods for the analysis of Quantitative Mass Spectrometry-based Proteomics data, especially for Liquid Chromatography-Mass Spectrometry (LC-MS) data. Indeed, the Ph.D. research project focused on understanding and overcoming the main problems related to bioinformatics for Quantitative Mass Spectrometry-based Proteomics and project suited software solutions to overcome them. In particular, efficient solutions to both data handling and quantification of profile LC-MS data were designed, implemented and validated.

This project focused on LC-MS data, which are deemed to be the only data source rich enough to carry out a meaningful Quantitative Mass Spectrometry-based Proteomics analysis. Data features pivotal for the design of the proposed solutions essentially are the 3D structure of LC-MS data and the high quality profile acquisition mode. In fact, LC-MS separates peptides in two dimensions (t, m/z) minimizing their overlap, and the profile acquisition mode enhances signal quantification.

In order to properly validate the developed algorithms, an appropriate dataset was used. It consists of LC-MS data from a controlled mixture of ICPL-labeled proteins with known ratios and triplicates. They were acquired in enhanced profile mode for survey scans to gain higher mass accuracy. Thus the quantitative informative content for this dataset is very high.

The proposed methods, assessed on this high quality dataset, demonstrated to outperform some well-known software solutions commonly used.

DATA HANDLING

Regarding the data handling issue a scalable 2D indexing approach was proposed. It is implemented through an R-tree-based data structure, called mzRTree, that relies on a sparse matrix representation of the dataset, which is appropriate for MS-based proteomics data. mzRTree is described in Chapter 6. mzRTree can be efficiently built and stored and ensures efficient 1D and 2D data access. Further results show that mzRTree requires the smallest hard disk space, data structure loading time and features an efficient creation time. Moreover, mzRTree is fairly scalable as regards access and data structure load time: as data density increases by a factor 10, the access time increases by a factor less than 3, while the load time is approximately constant. Experimental results and the R-tree structure scalability suggest that mzRTree is suitable for high density/large size proteomics data, such as 3D profile LC-MS data. Actually, these data are the only data source rich enough to perform a meaningful quantitative analysis. However, costs involved with profile data handling often outweigh their benefits. mzRTree could revert this relationship.

QUANTIFICATION

Quantification is one of the most important open issues in mass spectrometry-based proteomics. During this Ph.D. research, 3DSpectra, an innovative quantification algorithm for LC-MS labeled profile data was developed. It is described in Chapter 7. 3DSpectra accesses data using mzRTree and makes use of a priori information, provided by search engines, to quantify identified peptides, whose metadata are stored in a structured collection, the peptide library. 3DSpectra fits on peptide data the 3D isotopic distribution model shaped by a Gaussian Mixture Model (GMM) including a noise component, using the Expectation-Maximization (EM) approach. The EM starting parameters, i.e., Gaussians' centers and shapes, are retrieved by the metadata. Peaks' borders are recognized from the GMM iso-density curves and outlying data or data belonging to the noise component are discarded from analysis. 3DSpectra substantially improves quantification efficiency compared to ASAPRatio (MASPECTRAS implementation), and features the same good quantification accuracy, precision and reliability. Moreover, 3DSpectra achieves a significantly higher reproducibility of its peptide quantifications across experimental replicates.

FINAL REMARKS AND FUTURE WORK

In conclusion, during this PhD project 2 software solutions have been proposed to address the handling and quantification of Mass Spectrometry-based Quantitative Proteomics data: mzRTree and 3DSpectra, respectively. mzRTree allows efficient data access, storage and enables a computationally sustainable analysis of profile MS data. Regarding the quantification issue, 3DSpectra is a reliable and accurate quantification strategy for labeled LC-MS data, providing significantly wide and reproducible proteome coverage.

Future and ongoing research work is focused on further development of both the mzRTree data structure and 3DSpectra quantification software.

mzRTree capabilities will be exploited in order to help the community for storing and accessing MS data. Recently, mzRTree was proposed to the Proteomics Standards Initiative (PSI) community as a valuable computational support to existing standards. At the moment a project is under development regarding the possibility of making use of mzRTree to realize a new open data format compliant to computational requirements from data analysis. This research activity is carried on in collaboration with foreign researchers involved in the development of PSI data formats and ontologies.

3DSpectra will be applied to a broader range of proteomics MS-based data (e.g., Selected Reaction Monitoring data), considering also possible suitable modifications to the 3D model of the peptide distribution. Further work is then needed to facilitate the import of identification results and the export of quantification results, adding support for the PSI exchange data formats, i.e., mzIdentML and mzQuantML.

ACKNOWLEDGMENTS

First of all, I would like to acknowledge Juergen Hartler for following this work during the last 3 years, always providing significant feedback.

Special thanks to:

- my Professors, my supervisor Barbara Di Camillo, Ph.D. and Prof. Gianna M. Toffolo, who allowed me to pursue this research project;

- Prof. Andrea Pietracaprina and Francesco Silvestri, Ph.D. for the collaboration during mzRTree development and to Piero De Gol for his informatics support.

Thanks to Andreas Schmidt for giving me the needed information about data, to Elena Ceseracciu and Fabio Scarpa for their help with C++, to Marco Scutari for his consulting about 3DSpectra statistical validation, to Paolo Denti for providing with the Deming regression code.

This study was funded by the CARIPARO 2008/2010 "Systems biology approaches to infer gene regulation from gene and protein time series data" project.

Last, but not least, I would like to sincerely acknowledge all beloved people that supported me during this work: my family and my boyfriend, and my dear friends.

REFERENCES

[1] C.H. Ahrens, E. Brunner, and K. Basler, “Quantitative proteomics: a central technology for systems biology.,” *Journal of proteomics*, vol. 73, Feb. 2010, pp. 820-7.

[2] E.F. Petricoin, C. Belluco, R.P. Araujo, and L. a Liotta, “The blood peptidome: a higher dimension of information content for cancer biomarker discovery.,” *Nature reviews. Cancer*, vol. 6, Dec. 2006, pp. 961-7.

[3] R. Aebersold and M. Mann, “Mass spectrometry-based proteomics.,” *Nature*, vol. 422, Mar. 2003, pp. 198-207.

[4] O. Rinner, L.N. Mueller, M. Hubálek, M. Müller, M. Gstaiger, and R. Aebersold, “An integrated mass spectrometric and computational framework for the analysis of protein interaction networks.,” *Nature biotechnology*, vol. 25, Mar. 2007, pp. 345-52.

[5] E.D. Jeffery, “Method for Full Protein Sequence Mapping: LC-MS analysis,” *Nature Protocols*, 2007, pp. 1-8.

[6] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse, “Electrospray ionization-principles and practice,” *Mass Spectrometry Reviews*, vol. 9, 1990, pp. 37-70.

[7] J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse, “Electrospray ionization for mass spectrometry of large biomolecules.,” *Science*, vol. 246, 1989, pp. 64-71.

[8] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp, “Matrix-assisted ultraviolet laser desorption of non-volatile compounds,” *International Journal of Mass Spectrometry and Ion Processes*, vol. 78, 1987, pp. 53-68.

[9] M. Karas, D. Bachmann, and F. Hillenkamp, “Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules,” *Analytical Chemistry*, vol. 57, 1985, pp. 2935-2939.

[10] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, and T. Matsuo, “Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry,” *Rapid Communications in Mass Spectrometry*, vol. 2, 1988, pp. 151-153.

[11] G.W. Jr, L.H. Cazares, S.M. Leung, S. Nasim, B.L. Adam, T.T. Yip, P.F. Schellhammer, L. Gong, and A. Vlahou, “Proteinchip(R) surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures,” *Prostate Cancer Prostatic Dis*, vol. 2, 1999, pp. 264-276.

[12] W.C. Wiley and I.H. McLaren, “Time-of-Flight Mass Spectrometer with Improved Resolution,” *Review of Scientific Instruments*, vol. 26, 1955, p. 1150.

[13] A.G. Marshall, C.L. Hendrickson, and G.S. Jackson, “Fourier transform ion cyclotron resonance mass spectrometry: a primer.,” *Mass Spectrometry Reviews*, vol. 17, 1998, pp. 1-35.

[14] A. Makarov, “Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis,” *Analytical Chemistry*, vol. 72, 2000, pp. 1156-1162.

[15] Q. Hu, R.J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks, “The Orbitrap: a new mass spectrometer.,” *Journal of mass spectrometry JMS*, vol. 40, 2005, pp. 430-443.

[16] A. Makarov, “Theory and practice of the orbitrap mass analyzer Principle of Trapping in the Orbitrap,” *ASMS*, 2006, pp. 1-9.

[17] M. Scigelova and A. Makarov, “Orbitrap Mass Analyzer – Overview and Applications in Proteomics,” *Proteomics*, vol. 6, 2006, pp. 16-21.

[18] A. Makarov, E. Denisov, O. Lange, and S. Horning, “Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer.,” *Journal of the American Society for Mass Spectrometry*, vol. 17, 2006, pp. 977-982.

[19] R.A. Yost and C.G. Enke, “Selected ion fragmentation with a tandem quadrupole mass spectrometer,” *Journal of the American Chemical Society*, vol. 100, 1978, pp. 2274-2275.

[20] R.E. March, “An Introduction to Quadrupole Ion Trap Mass Spectrometry,” *Journal of Mass Spectrometry*, vol. 32, 1997, pp. 351-369.

[21] W. Paul and H. Steinwedel, “Ein neues Massenspektrometer ohne Magnetfeld,” *Zeitschrift für Naturforschung*, vol. 8, 1953, pp. 448-450.

[22] L. Linsen, J. Löcherbach, M. Berth, D. Becher, and J. Bernhardt, “Visual analysis of gel-free proteome data.,” *IEEE transactions on visualization and computer graphics*, vol. 12, 2006, pp. 497-508.

[23] L. Martens, A.I. Nesvizhskii, H. Hermjakob, M. Adamski, G.S. Omenn, J. Vandekerckhove, and K. Gevaert, “Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories.,” *Proteomics*, vol. 5, Aug. 2005, pp. 3501-5.

[24] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold, “Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.,” *Nature Biotechnology*, vol. 17, 1999, pp. 994-999.

[25] J.D. Jaffe, D.R. Mani, K.C. Leptos, G.M. Church, M. a Gillette, and S. a Carr, “PEPPeR, a platform for experimental proteomic pattern recognition.,” *Molecular & cellular proteomics : MCP*, vol. 5, Oct. 2006, pp. 1927-41.

[26] M. Mann, “Functional and quantitative proteomics using SILAC,” *Nature Reviews Molecular Cell Biology*, vol. 7, 2006, pp. 952-958.

[27] T. Geiger, J. Cox, P. Ostasiewicz, J.R. Wisniewski, and M. Mann, “Super-SILAC mix for quantitative proteomics of human tumor tissue.,” *Nature Methods*, vol. 7, 2010, pp. 383-385.

[28] S.D. Patterson and R.H. Aebersold, “Proteomics: the first decade and beyond.,” *Nature genetics*, vol. 33 Suppl, Mar. 2003, pp. 311-23.

[29] L. Choe, M. D’Ascenzo, N.R. Relkin, D. Pappin, P. Ross, B. Williamson, S. Guertin, P. Pribil, and K.H. Lee, “8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer’s disease.,” *Proteomics*, vol. 7, 2007, pp. 3651-3660.

[30] S.-E. Ong, L.J. Foster, and M. Mann, “Mass spectrometric-based approaches in quantitative proteomics,” *October*, vol. 29, 2003, pp. 124-130.

[31] H. Liu, R.G. Sadygov, and J.R. Yates, “A model for random sampling and estimation of relative protein abundance in shotgun proteomics.,” *Analytical Chemistry*, vol. 76, 2004, pp. 4193-4201.

[32] J. Listgarten and A. Emili, “Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry.,” *Molecular & cellular proteomics : MCP*, vol. 4, Apr. 2005, pp. 419-34.

[33] R. Kiyonami, A. Schoen, A. Prakash, S. Peterman, V. Zabrouskov, P. Picotti, R. Aebersold, A. Huhmer, and B. Domon, “Increased selectivity, analytical precision, and throughput in targeted proteomics.,” *Molecular cellular proteomics MCP*, 2010, p. in press.

[34] S. Elschenbroich and T. Kislinger, “Targeted proteomics by selected reaction monitoring mass spectrometry: applications to systems biology and biomarker discovery,” *Molecular Biosystems*, 2010.

[35] R. Kiyonami and B. Domon, “Selected reaction monitoring applied to quantitative proteomics.,” *Methods In Molecular Biology Clifton Nj*, vol. 658, 2010, pp. 155-166.

[36] V. Lange, P. Picotti, B. Domon, and R. Aebersold, “Selected reaction monitoring for quantitative proteomics: a tutorial,” *Molecular Systems Biology*, vol. 4, 2008, p. 222.

[37] M. Vaudel, A. Sickmann, and L. Martens, “Peptide and protein quantification: a map of the minefield.,” *Proteomics*, vol. 10, Feb. 2010, pp. 650-70.

[38] A. SAVITZKY and M.J.E. GOLAY, “Smoothing and differentiation of data by simplified least squares procedures.,” *Analytical Chemistry*, vol. 36, 1964, pp. 1627-1639.

[39] C.E. Bakalarski, J.E. Elias, J. Villén, W. Haas, S. a Gerber, P. a Everley, and S.P. Gygi, “The impact of peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses.,” *Journal of proteome research*, vol. 7, Nov. 2008, pp. 4756-65.

[40] W. Wang, H. Zhou, H. Lin, S. Roy, T. a Shaler, L.R. Hill, S. Norton, P. Kumar, M. Anderle, and C.H. Becker, “Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards.,” *Analytical chemistry*, vol. 75, Sep. 2003, pp. 4818-26.

[41] D. Radulovic, S. Jelveh, S. Ryu, T.G. Hamilton, E. Foss, Y. Mao, and A. Emili, “Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry.,” *Molecular & cellular proteomics : MCP*, vol. 3, Oct. 2004, pp. 984-97.

[42] K.C. Leptos, D. a Sarracino, J.D. Jaffe, B. Krastins, and G.M. Church, “MapQuant: open-source software for large-scale protein quantification.,” *Proteomics*, vol. 6, Mar. 2006, pp. 1770-82.

[43] 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, Jun. 1991, pp. 583-598.

[44] P. Wang, H. Tang, M.P. Fitzgibbon, M. McIntosh, M. Coram, H. Zhang, E. Yi, and R. Aebersold, “A statistical method for chromatographic alignment of LC-MS data.,” *Biostatistics (Oxford, England)*, vol. 8, Apr. 2007, pp. 357-67.

[45] P. Wang, H. Tang, H. Zhang, J. Whiteaker, A.G. Paulovich, and M. Mcintosh, “Normalization regarding non-random missing values in high-throughput mass spectrometry data.,” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 326, Jan. 2006, pp. 315-26.

[46] J. Listgarten, R.M. Neal, S.T. Roweis, and A. Emili, “Multiple Alignment of Continuous Time Series,” *Constraints*, vol. 17, 2005, pp. 817-824.

[47] J. Quackenbush, “Microarray data normalization and transformation.,” *Nature Genetics*, vol. 32 Suppl, 2002, pp. 496-501.

[48] C.F. Taylor, N.W. Paton, K.S. Lilley, P.-A. Binz, R.K. Julian, A.R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E.W. Deutsch, M.J. Dunn, A.J.R. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. a Neubert, S.D. Patterson, P. Ping, S.L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T.M. Vondriska, J.P. Whitelegge, M.R. Wilkins, I. Xenarios, J.R. Yates, and H. Hermjakob, “The minimum information about a proteomics experiment (MIAPE).,” *Nature biotechnology*, vol. 25, Aug. 2007, pp. 887-93.

[49] P.-A. Binz, R. Barkovich, R.C. Beavis, D. Creasy, D.M. Horn, R.K. Julian, S.L. Seymour, C.F. Taylor, and Y. Vandenbrouck, “Guidelines for reporting the use of mass spectrometry informatics in proteomics.,” *Nature Biotechnology*, vol. 26, 2008, p. 862.

[50] K. Garwood, T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C.F. Taylor, K. Carroll, C. Evans, A.D. Whetton, S. Hart, D. Stead, Z. Yin, A.J. Brown, A. Hesketh, K. Chater, L. Hansson, M. Mewissen, P. Ghazal, J. Howard, K.S. Lilley, S.J. Gaskell, A. Brass, S.J. Hubbard, S.G. Oliver, and N.W. Paton, “PEDRo: A database for storing, searching and disseminating experimental proteomics data,” *BMC Genomics*, vol. 5, 2004, p. 68.

[51] L. Martens, H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove, and R. Apweiler, “PRIDE: the proteomics identifications database.,” *Proteomics*, vol. 5, Aug. 2005, pp. 3537-45.

[52] H. Hermjakob and R. Apweiler, “The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible.,” *Expert Review Of Proteomics*, vol. 3, 2006, pp. 1-3.

[53] F. Desiere, E.W. Deutsch, A.I. Nesvizhskii, P. Mallick, N.L. King, J.K. Eng, A. Aderem, R. Boyle, E. Brunner, S. Donohoe, N. Fausto, E. Hafen, L. Hood, M.G. Katze, K.A. Kennedy, F. Kregenow, H. Lee, B. Lin, D. Martin, J.A. Ranish, D.J. Rawlings, L.E. Samelson, Y. Shiio, J.D. Watts, B. Wollscheid, M.E. Wright, W. Yan, L. Yang, E.C. Yi, H. Zhang, and R. Aebersold, “Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry,” *Genome Biology*, vol. 6, 2005, p. R9.

[54] E.W. Deutsch, J.K. Eng, H. Zhang, N.L. King, A.I. Nesvizhskii, B. Lin, H. Lee, E.C. Yi, R. Ossola, and R. Aebersold, “Human Plasma PeptideAtlas.,” *Proteomics*, vol. 5, 2005, pp. 3497-3500.

[55] E.W. Deutsch, H. Lam, and R. Aebersold, “PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows.,” *EMBO Reports*, vol. 9, 2008, pp. 429-434.

[56] P. Picotti, H. Lam, D. Campbell, E.W. Deutsch, H. Mirzaei, J. Ranish, B. Domon, and R. Aebersold, “A database of mass spectrometric assays for the yeast proteome.,” *Nature Methods*, vol. 5, 2008, pp. 913-914.

[57] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.," *Nature Genetics*, vol. 29, 2001, pp. 365-371.

[58] S. Orchard, L. Montechi-Palazzi, E.W. Deutsch, P.-A. Binz, A.R. Jones, N. Paton, A. Pizarro, D.M. Creasy, J. Wojcik, and H. Hermjakob, "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 Supérieure (ENS), Lyon, France.," *Proteomics*, vol. 7, Oct. 2007, pp. 3436-40.

[59] P.G. a Pedrioli, J.K. Eng, R. Hubley, M. Vogelzang, E.W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R.H. Angeletti, R. Apweiler, K. Cheung, C.E. Costello, H. Hermjakob, S. Huang, R.K. Julian, E. Kapp, M.E. McComb, S.G. Oliver, G. Omenn, N.W. Paton, R. Simpson, R. Smith, C.F. Taylor, W. Zhu, and R. Aebersold, "A common open representation of mass spectrometry data and its application to proteomics research.," *Nature biotechnology*, vol. 22, Nov. 2004, pp. 1459-66.

[60] T. Bray, "Extensible Markup Language (XML)," *W3C recommendation*, vol. 6, 2000, pp. 274-276.

[61] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W.H. Tang, A. Rompp, S. Neumann, A.D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, and E.W. Deutsch, "mzML - a Community Standard for Mass Spectrometry Data.," *Molecular & cellular proteomics : MCP*, Aug. 2010, pp. 2010-2010.

[62] S. Josefsson, "The Base16, Base32, and Base64 Data Encodings," 2006.

[63] S.M. Lin, L. Zhu, A.Q. Winter, M. Sasinowski, and W.A. Kibbe, "What is mzXML good for?," *Expert Review Of Proteomics*, vol. 2, 2005, pp. 839-845.

[64] S. Purvine, J.-T. Eppel, E.C. Yi, and D.R. Goodlett, "Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer.," *Proteomics*, vol. 3, 2003, pp. 847-850.

[65] T. Iupac, I. Nomenclature, B. Compounds, R. Nomenclature, S. Designations, and S. Topic, "IUPAC Nomenclature of Organic Compounds," 1993, pp. 1-5.

[66] J. Hartler, G.G. Thallinger, G. Stocker, A. Sturn, T.R. Burkard, E. Körner, R. Rader, A. Schmidt, K. Mechtler, and Z. Trajanoski, "MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data.," *BMC bioinformatics*, vol. 8, Jan. 2007, p. 197.

[67] C.U. Mohien, J. Hartler, F. Breitwieser, U. Rix, L.R. Rix, G.E. Winter, G.G. Thallinger, K.L. Bennett, G. Superti-Furga, Z. Trajanoski, and J. Colinge, "MASPECTRAS 2: An integration and analysis platform for proteomic data.," *Proteomics*, vol. 10, Jul. 2010, pp. 2719-22.

[68] Z. Khan, J.S. Bloom, B. a Garcia, M. Singh, and L. Kruglyak, "Protein quantification across hundreds of experimental conditions.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, Sep. 2009, pp. 15544-8.

[69] L.N. Mueller, M.-Y. Brusniak, D.R. Mani, and R. Aebersold, "An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data.," *Journal of proteome research*, vol. 7, Jan. 2008, pp. 51-61.

[70] R. Matthiesen, "Methods, algorithms and tools in computational proteomics: a practical point of view.," *Proteomics*, vol. 7, Aug. 2007, pp. 2815-32.

[71] P.M. Palagi, P. Hernandez, D. Walther, and R.D. Appel, "Proteome informatics I: bioinformatics tools for processing experimental data.," *Proteomics*, vol. 6, Oct. 2006, pp. 5435-44.

[72] X.-J. Li, H. Zhang, J. a Ranish, and R. Aebersold, "Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry.," *Analytical chemistry*, vol. 75, Dec. 2003, pp. 6648-57.

[73] E.W. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, H. Lam, N. Tasman, Z. Sun, E. Nilsson, B. Pratt, B. Prazen, J.K. Eng, D.B. Martin, A.I. Nesvizhskii, and R. Aebersold, "A guided tour of the Trans-Proteomic Pipeline.," *Proteomics*, vol. 10, Mar. 2010, pp. 1150-9.

[74] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.," *Nature biotechnology*, vol. 26, Dec. 2008, pp. 1367-72.

[75] K.-Y. Leung, P. Lescuyer, J. Campbell, H.L. Byers, L. Allard, J.-C. Sanchez, and M. a Ward, "A novel strategy using MASCOT Distiller for analysis of cleavable isotope-coded affinity tag data to quantify protein changes in plasma.," *Proteomics*, vol. 5, Aug. 2005, pp. 3040-4.

[76] N. Colaert, K. Helsens, F. Impens, J. Vandekerckhove, and K. Gevaert, "Rover: a tool to visualize and validate quantitative proteomics data from different sources.," *Proteomics*, vol. 10, Mar. 2010, pp. 1226-9.

[77] X.-J. Li, H. Zhang, J. a Ranish, and R. Aebersold, "Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry.," *Analytical chemistry*, vol. 75, Dec. 2003, pp. 6648-57.

[78] J.K. Eng, B. Fischer, J. Grossmann, and M.J. Maccoss, "A fast SEQUEST cross correlation algorithm.," *Journal of proteome research*, vol. 7, Oct. 2008, pp. 4598-602.

[79] M. Brosch, L. Yu, T. Hubbard, and J. Choudhary, "Accurate and sensitive peptide identification with Mascot Percolator.," *Journal of proteome research*, vol. 8, Jun. 2009, pp. 3176-81.

[80] A.I. Nesvizhskii and R. Aebersold, "Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS.," *Drug Discovery Today*, vol. 9, 2004, pp. 173-181.

[81] P. Mortensen, J.W. Gouw, J.V. Olsen, S.-E. Ong, K.T.G. Rigbolt, J. Bunkenborg, J. Cox, L.J. Foster, A.J.R. Heck, B. Blagoev, J.S. Andersen, and M. Mann, "MSQuant, an open source platform for mass spectrometry-based quantitative proteomics.," *Journal of proteome research*, vol. 9, Jan. 2010, pp. 393-403.

[82] P. Mortensen, J.W. Gouw, J.V. Olsen, S.-E. Ong, K.T.G. Rigbolt, J. Bunkenborg, J. Cox, L.J. Foster, A.J.R. Heck, B. Blagoev, J.S. Andersen, and M. Mann, "MSQuant, an open source platform for mass spectrometry-based quantitative proteomics.," *Journal of proteome research*, vol. 9, Jan. 2010, pp. 393-403.

[83] S.K. Park, J.D. Venable, T. Xu, and J.R. Yates, "A quantitative analysis software tool for mass spectrometry-based proteomics.," *Nature methods*, vol. 5, Apr. 2008, pp. 319-22.

[84] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher, "OpenMS – An open-source software framework for mass spectrometry," *BMC Bioinformatics*, vol. 9, 2008, p. 163.

[85] O. Kohlbacher, K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm, “TOPP--the OpenMS proteomics pipeline.,” *Bioinformatics (Oxford, England)*, vol. 23, Jan. 2007, pp. e191-7.

[86] O. Schulz-trieglaff, R. Hussong, and C. Gr, “A Fast and Accurate Algorithm for the Quantification of Peptides from Mass Spectrometry Data,” *Methods*.

[87] C. Gröpl, E. Lange, K. Reinert, O. Kohlbacher, M. Sturm, C.G. Huber, B.M. Mayr, and C.L. Klein, “Algorithms for the Automated Absolute Quantification of Diagnostic Markers in Complex Proteomics Samples,” *Computational Life Sciences*, 2005, pp. 151-162.

[88] B.M. Mayr, O. Kohlbacher, K. Reinert, M. Sturm, C. Gröpl, E. Lange, C. Klein, and C.G. Huber, “Absolute myoglobin quantitation in serum by combining two-dimensional liquid chromatography-electrospray ionization mass spectrometry and novel data analysis algorithms.,” *Journal of Proteome Research*, vol. 5, 2006, pp. 414-421.

[89] A. Guttman, “R-trees: a dynamic index structure for spatial searching,” *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, B. Yormack, ed., ACM New York, NY, USA, 1984, pp. 47-57.

[90] J.S. Vitter, “External Memory Algorithms and Data Structures: Dealing with Massive Data,” *ACM Computing Surveys*, vol. 33, 2001, pp. 209-271.

[91] O. Schulz-Trieglaff, N. Pfeifer, C. Gröpl, O. Kohlbacher, and K. Reinert, “LC-MSsim--a simulation software for liquid chromatography mass spectrometry data.,” *BMC bioinformatics*, vol. 9, Jan. 2008, p. 423.

[92] S. Nasso, F. Silvestri, F. Tisiot, B. Di Camillo, A. Pietracaprina, and G.M. Toffolo, “An optimized data structure for high-throughput 3D proteomics data: mzRTree.,” *Journal of proteomics*, vol. 73, Apr. 2010, pp. 1176-82.

[93] R.D. Bjornson, N.J. Carriero, C. Colangelo, M. Shifman, K.-H. Cheung, P.L. Miller, and K. Williams, “X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers.,” *Journal of Proteome Research*, vol. 7, 2008, pp. 293-299.

[94] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley-Interscience, 2000.

[95] J.D. Jaffe, D.R. Mani, K.C. Leptos, G.M. Church, M. a Gillette, and S. a Carr, “PEPPeR, a platform for experimental proteomic pattern recognition.,” *Molecular & cellular proteomics : MCP*, vol. 5, Oct. 2006, pp. 1927-41.

[96] J.A. Bilmes, “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models,” *International Computer Science Institute*, vol. 4, 1998, p. 15.

[97] C. Fraley, “Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering,” *Journal of Classification*, vol. 181, 2007, pp. 155-181.

[98] P. Paalanen, J. Kamarainen, J. Ilonen, and H. Kalviainen, “Feature representation and discrimination based on Gaussian mixture model probability densities—Practices and algorithms,” *Pattern Recognition*, vol. 39, Jul. 2006, pp. 1346-1358.

[99] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. McIntosh, “A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS.,” *Bioinformatics (Oxford, England)*, vol. 22, Aug. 2006, pp. 1902-9.

[100] J. Hartler, G.G. Thallinger, G. Stocker, A. Sturn, T.R. Burkard, E. Körner, R. Rader, A. Schmidt, K. Mechtler, and Z. Trajanoski, “MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data.,” *BMC bioinformatics*, vol. 8, Jan. 2007, p. 197.

[101] W.E. Deming, “Statistical adjustment of data,” *New York USA John Wiley and Sons*, 1943.

[102] K. Linnet, “Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies.,” *Clinical Chemistry*, vol. 44, 1998, pp. 1024-1031.

[103] D.S. Smith, M. Pourfarzaneh, and R.S. Kamel, “Linear regression analysis by Deming’s method.,” *Clinical Chemistry*, vol. 26, 1980, pp. 1105-1106.

[104] S. Orchard, J.-P. Albar, E.W. Deutsch, M. Eisenacher, P.-A. Binz, and H. Hermjakob, “implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada.,” *Proteomics*, vol. 10, 2010, pp. 1895-1898.

[105] S. Orchard, A. Jones, J.-P. Albar, S.Y. Cho, K.-H. Kwon, C. Lee, and H. Hermjakob, “Tackling quantitation: a report on the annual Spring Workshop of the HUPO-PSI 28-30 March 2010, Seoul, South Korea.,” *Proteomics*, vol. 10, 2010, pp. 3062-3066.

APPENDIX A

Regarding the PSI data formats, here below their copyright notice is reported.

“**Intellectual Property Statement**

The PSI takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the PSI Secretariat.

The PSI invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this recommendation. Please address the information to the PSI Executive Director (see contacts information at PSI website).

Full Copyright Notice

Copyright (C) Proteomics Standards Initiative (2006). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the PSI or other organizations, except as needed for the purpose of developing Proteomics Recommendations in which case the procedures for copyrights defined in the PSI Document process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the PSI or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE PROTEOMICS STANDARDS INITIATIVE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.”

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

HPLC	High Performance Liquid Chromatography
NSI	Nano Spray Ionization
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
LC-MS	Liquid Chromatography - Mass Spectrometry
m/z	Mass-To-Charge Ratio
3D	3-Dimensional
GMM	Gaussian Mixture Model
ML	Maximum Likelihood
EM	Expectation Maximization
ICPL	Isotope - Coded Protein Labels
JRAP	Java Random Access Library
FMM	Finite Mixture Modeling
PDF	Probability Density Function
MEX	MATLAB Executable
WLLS	Weighted Linear Least Squares
VUC	Volume Under The Curve
LTQ	Linear Trap Quadrupole
ITMS	Ion Trap Mass Spectrometer
SD	Standard Deviation
CV%	Percentage Coefficient Of Variability
R^2	Pearson’s correlation coefficient squared value
RMSE	Root Mean Squared Error
F	Fisher-Snedecor
SRM	Selected Reaction Monitoring