Computational Methods for Quantitative Peptide Mass Spectrometry

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

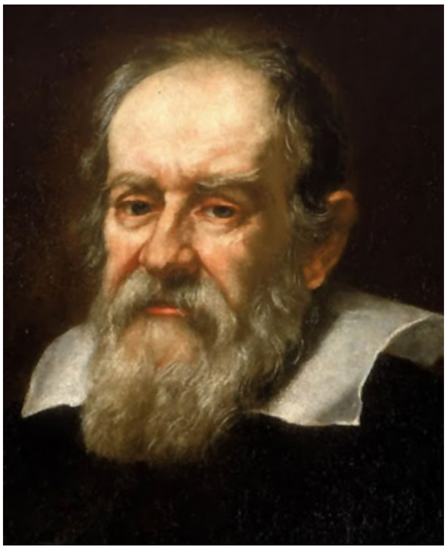
Ole Schulz-Trieglaff

Berlin
October 2008

Gutachter:

Professor Dr. Knut Reinert, Freie Universität Berlin
Professor Dr. Oliver Kohlbacher, Eberhard Karls Universität Tübingen

Datum der Disputation: 11. Juni 2009



If the data refute the theory, the data are probably wrong.

Galileo Galilei (1564 - 1642), Italian physicist and mathematician.

Abstract

This thesis presents algorithms for the analysis of liquid chromatography-mass spectrometry (LC-MS) data. Mass spectrometry is a technology that can be used to determine the identities and abundances of the compounds in complex samples. In combination with liquid chromatography, it has become a popular method in the field of proteomics, the large-scale study of proteins and peptides in living systems. This area of research has gained a lot of interest in recent years since proteins control fundamental reactions in the cell. Consequently, a deeper knowledge of their function is expected to be crucial for the development of new drugs and the cure of diseases.

The data sets obtained from an LC-MS experiment are large and highly complex. The outcome of such an experiment is called an LC-MS map. The map is a collection of mass spectra. They contain, among the signals of interest, a high amount of noise and other disturbances. That is why algorithms for the low-level processing of LC-MS data are becoming increasingly important. These algorithms are the focus of this text.

Our novel contributions are threefold: first, we introduce SweepWavelet, an algorithm for the efficient detection and quantification of peptides from LC-MS data. The quantification of proteins and peptides using mass spectrometry is of high interest for biomedical research but also for the pharmaceutical industry since it is usually among the first steps in an LC-MS data analysis pipeline and all subsequent steps depend on its quality. Our approach was among the first to address this problem in a sound computational framework. It consists of three steps: first, we apply a tailored wavelet function that filters mass spectra for the isotope peaks of peptides. Second, we use a method inspired by the sweep-line paradigm which makes use of the redundant information in LC-MS data to determine mass, charge, retention time and abundance of all peptides. Finally, we apply a flexible peptide signal model to filter the extracted signals for false positives.

The second part of this thesis deals with the benchmarking of LC-MS signal detection algorithms. This is a non-trivial task since it is difficult to establish a ground truth using real world samples: which sample compounds become visible in an LC-MS data set is not known in advance. To this end, we use annotated data and simulations to assess the performance of currently available algorithms. To simulate benchmark data, we developed a simulation software called LC-MSsim. It incorporates computational models for retention time prediction, peptide detectability, isotope pattern and elution peaks. Using this software, we can simulate all steps in an LC-MS experiment and obtain a list with the positions, charges and abundances of all peptide signals contained in the resulting LC-MS map. This gives us a ground truth against which we can match the results of a signal detection algorithm. In this thesis, we use it for the benchmarking of quantification algorithms but its scope is wider and it can also be used to evaluate other algorithms. To our knowledge, LC-MSsim is the first software that can simulate the full LC-MS data acquisition process.

The third contribution of this thesis is a statistical framework for the quality assessment of quantitative LC-MS experiments. Whereas quality assessment and control are already widespread in the field of gene expression analysis, our work is the first to address this problem for LC-MS data. We use methods from robust statistics to detect outlier LC-MS maps in large-scale quantitative experiments. Our approach introduces the notion of quality descriptors to derive an abstract representation of an LC-MS map and applies a robust principal component analysis based on projection pursuit. We show that it is sensible to use robust statistics for this problem and evaluate our method on simulated maps and on data from three real-world LC-MS studies.

Zusammenfassung

Das Thema dieser Arbeit sind Algorithmen für die Analyse vom Flüssigchromatographie-Massenspektrometrie (LC-MS) Daten. Das Ergebnis eines LC-MS Experiments wird LC-MS Map genannt. Die Map ist eine Gruppe von Massenspektren. Mit Hilfe der Massenspektrometrie lassen sich komplexe biologische Proben auf ihre Zusammensetzung untersuchen. In Kombination mit Flüssigchromatographie ist sie zu einem wichtigen Werkzeug in der Proteomik geworden. Die Proteomik umfasst die Erforschung des Proteoms, das heißt der Gesamtheit aller in einer Probe vorhandenen Proteine und Peptide. Proteomik als wissenschaftliche Disziplin ist den letzten Jahren sehr populär geworden, da Proteine essentielle Reaktionen in der Zelle steuern und als wichtige Angriffspunkte für die Diagnose und Heilung von Krankheiten gelten.

Diese Arbeit enthält drei neue wissenschaftliche Beiträge. Der erste ist SweepWavelet, ein Algorithmus zur Quantifizierung von Peptiden aus LC-MS Daten. Die akkurate Quantifizierung von Peptiden und Proteinen ist ein wichtiges Thema in der biomedizinischen Forschung, da sie der erste Schritt in der rechnergestützten Analyse von LC-MS Daten ist. Alle weiteren Schritte hängen von einer präzisen und zuverlässigen Quantifizierung ab. Im Gegensatz zu bestehenden Verfahren ist unser Algorithmus flexibel, schnell und kann leicht an Datensätze von verschiedenen LC-MS Instrumenten angepasst werden. Unser Algorithmus besteht aus drei Schritten: wir verwenden eine Wavelet Funktion um Peptidsignale aus den LC-MS Daten herauszufiltern und Hintergrundrauschen zu unterdrücken. Danach benutzen wir die sweep-line Methode aus der algorithmischen Geometrie um effizient die Position der Peptidsignale im LC-MS Datensatz zu bestimmen und ihre Abundanz zu schätzen. Im dritten Teil des Algorithmus verwenden wir ein flexibles Modell von LC-MS Peptidsignalen um falsch positive Signale zu entfernen.

Der zweite Teil dieser Arbeit widmet sich dem Vergleich von Algorithmen zur Peptidsignalerkennung und -quantifizierung. Dies ist ein schwieriges Unterfangen, da man in echten LC-MS Experimenten im Voraus nicht mit Sicherheit bestimmen kann, welche Substanzen in der LC-MS Map als Signale auftreten und welche nicht. Deshalb sind die Resultate von Algorithmen oft schwer zu beurteilen. Wir führen Vergleiche auf echten und simulierten Daten durch. Zu diesem Zweck haben wir eine Simulationssoftware für LC-MS Experimente entwickelt. Diese Software, LC-MSsim, simuliert alle Teilschritte eines LC-MS Experiments, u.a. die Vorhersage von Retentionszeiten, Elutionsprofile und Hintergrundrauschen in den Spektren. Das Ergebnis einer Simulation ist ein künstlicher LC-MS Datensatz mit einer Liste der Positionen, Ladungen und Intensitäten aller Peptidsignale. Wir verwenden den Simulator um verschiedene Algorithmen zur Peptidquantifizierung zu vergleichen. Die Software ist unter einer Open Source Lizenz frei verfügbar. LC-MSsim ist die erste frei verfügbare Software, welche vollständige LC-MS Datensätze inklusive der wichtigsten experimentellen Schritte simulieren kann.

Der dritte Beitrag dieser Arbeit ist eine neue statistische Methode zur Erkennung von Ausreißern bzw. Datensätzen schlechter Qualität in LC-MS Studien. Diese Methode basiert auf einer projection pursuit Version der Hauptkomponentenanalyse. Der Vorteil des projection pursuit Ansatzes ist seine Robustheit gegenüber Ausreißern. In anderen wissenschaftlichen Gebieten, wie z.B. der Genexpressionsanalyse, sind Methoden zur Qualitätskontrolle weit verbreitet. Unsere Methode gehört jedoch zu den ersten die sich der Qualitätskontrolle in LC-MS gestützten Studien widmet. Gerade in Hochdurchsatzexperimenten ist es äußerst wichtig, schlechte Messungen schnell entfernen zu können, um aussagekräftige Ergebnisse zu erhalten. Wir evaluieren unsere Methode auf simulierten und echten Daten und zeigen, dass wir Ausreißer schnell und präzise identifizieren können.

Acknowledgments

I am highly indebted to my advisor Knut Reinert. He suggested the topic of this thesis and provided valuable feedback and scientific input for this work. I also spent a lot of time in discussions with Clemens Gröpl. Clemens is a constant source of creative ideas and he taught me a lot about mass spectrometry and bioinformatics in general. It was a pleasure to work with both of them.

I also would like to acknowledge my scientific collaborators: chapter 3 of this thesis results from a collaboration with Andreas Hildebrandt and Rene Hussong at the University of Saarbrücken. They introduced me to wavelet theory and provided the draft implementation of the isotope wavelet that I used in this work. The second part of chapter 4 is joint work with Nico Pfeifer and Oliver Kohlbacher at the University of Tübingen. Oliver also acts as the external referee for this thesis and I would to thank both of them for their collaboration and scientific input. Chapter 5 of this thesis are a result of a joint project with Hartmut Schlüter (Charite University Hospital Berlin) and Egidijus Machtejevas (Merck KGaA, Darmstadt). Hartmut and Egidijus provided the LC-MS maps that I used in this chapter and both answered my questions about chemistry and mass spectrometry patiently.

During my graduate studies, I spent three months at the research group of Terry Speed at the Walter and Elisa Hall Institute (WEHI) in Melbourne, Australia. Terry Speed is one of the most energetic persons I ever met, he has a great working style and clear way of thinking. I also would like to thank all other members of the Bioinformatics division at the WEHI, namely Alex, Belinda, Gordon, Greg and Mark who taught me a lot about statistics and data analysis and made my stay in Melbourne very enjoyable. Chapter 5 is the result of this stay.

I also would like to thank all members of the Bioinformatics and Software Engineering research groups at the Department of Computer Science at the Free University Berlin who suffered from my bad jokes for three years and were still able to keep up a good mood. In particular, I am indebted to Chris, Clemens, David, Gesine, Knut and Markus who proofread parts of this thesis and provided valuable feedback. It is much appreciated.

The International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC) and the Max Planck Institute for Molecular Genetics provided the funding for my PhD thesis and a unique research environment. Especially Hannes Luz, the coordinator of the IMPRS-CBSC, was always willing to help in a unbureaucratic manner.

In this thesis, I made heavy use of the OpenMS software library. I also integrated some of the software developed during this thesis into OpenMS. I would like to thank all OpenMS developers for putting so much effort into the development of this software.

Last but not least I would like to thank my family and Lucia Castellanos Perez-Bolde for their support during these three years.

Contents

Chapter 1

Introduction

Synopsis: *We motivate the topic of this thesis, the development and application of algorithms for mass spectrometry data analysis. Since we focus on the analysis of proteomics data, we give a brief introduction to molecular biology and protein chemistry.*

1.1 Motivation

After the completion of the Human Genome Project, research on proteins, the gene products, is the next consequent step. Proteins are key players in the cell: they are involved in signaling pathways, cell death, muscle contraction and other crucial body functions. In addition, proteins circulate through the body and interact with most organs. Therefore, proteins are also thought to be excellent biomarker for clinical studies. Accordingly, *proteomics* as a scientific discipline has received increased attention in recent years. In sum, this field aims at a complete understanding of the protein content of a biological sample, the study of sequence, structure and abundance of all proteins in a cell or body fluid at a given time and condition (for instance in a comparison of disease versus healthy probands or cells in different stages of growth). Proteomics meets growing interest in basic research but also in the pharmaceutical industry. On the one hand, scientists in academia try to understand cellular functions at a fundamental level, on the other hand, pharmaceutical companies are interested in better knowledge on drug metabolism or changes in protein levels in the presence of new drugs.

A typical proteomics experiment requires the analysis of highly complex biological samples, i.e., a sample containing many proteins at varying concentrations. In most cases, proteins will only be one class of molecules among many other sample compounds. Established technologies have problems to cope with this complexity. To give an example, two-dimensional gel electrophoresis is a classical method to analyze mixtures of proteins. It involves the separation of a protein mixture by two orthogonal properties, usually isoelectric point and molecular weight. Even if this method allows the simultaneous analysis of several thousand of proteins at the same time, the sample preparation is laborious, the electrophoresis is not well reproducible and has a relatively small dynamic range. Besides, it is also difficult to automate for high-throughput experiments. Other methods include enzyme-linked immunosorbent assays (ELISA) or Western blots, but they share most of the disadvantages of gel electrophoresis.

For these reasons, mass spectrometry-based methods have become increasingly popular. This technology can quantify and sequence proteins in complex samples and in a high-throughput manner. This thesis deals with data from a specific type of mass spectrometry experiment in which a chromatographic column is coupled to the mass spectrometer. This is called a liquid chromatography-mass spectrometry (LC-MS) experiment. We will give details of this technology in Chap. 2. Its applications to proteomics are numerous and include the study of protein-protein interactions (Nelson *et al.*, 2002), the identification of proteins in complex mixtures using database searching (Washburn *et al.*, 2001; Wolters *et al.*, 2001), de-novo sequencing of proteins (Bandeira *et al.*, 2007a,b), improving genome annotations (Jaffe *et al.*, 2004; Tanner *et al.*, 2007), the generation of protein expression data for different species (Griffin *et al.*, 2002; Lasonder *et al.*, 2002) and the identification of post-translational modifications (Schweppe *et al.*, 2003; Zhou *et al.*, 2001).

At the same time, quality but also quantity of the data generated from mass spectrometry experiments has increased dramatically. New instruments allow highly accurate mass measurements (Smith *et al.*, 2002), but also generate vast amounts of data. To give an example, a complex sample measured on a modern high-resolution mass spectrometer coupled to a chromatographic column results in a file of 1-2 GB size. In a typical study, we are dealing with samples from several time points or conditions and to make statistically sound statements, we need to measure each sample in technical replicates. Data sizes increase even more if methods such as multi-dimensional chromatography or MS/MS fragmentation are applied. These problems have lead the scientific community to the conclusion that efficient algorithms are crucial

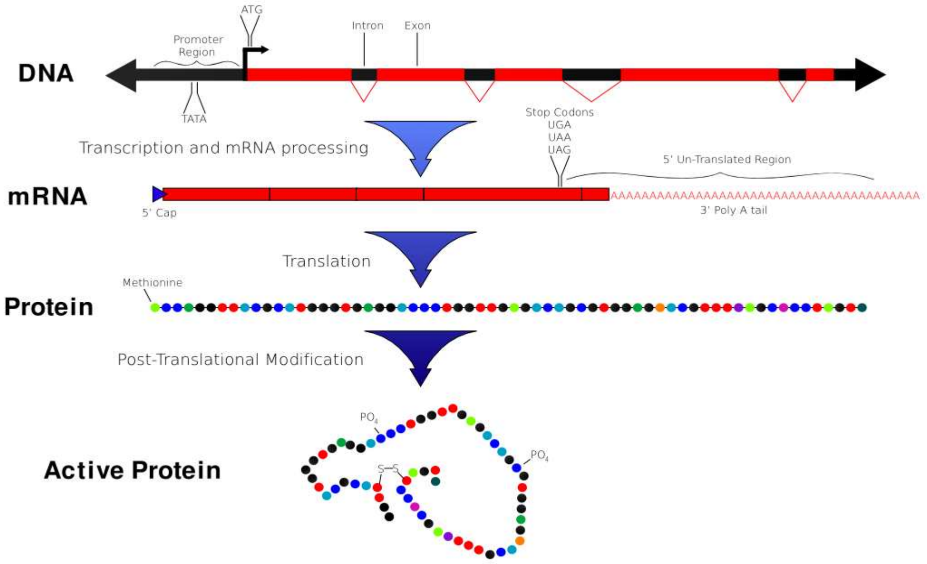


Figure 1.1: The central dogma of molecular biology (illustration modified from the corresponding wikipedia entry, http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology, accessed August 2008.). The dogma states that the transfer of information in the human cell goes from DNA over RNA to proteins. Proteins were assumed to be the only molecules with a regulatory function. This idea has been challenged and it is now widely accepted that other molecules, namely RNA, have also important regulatory function (Gilbert, 1986). But the dogma is still considered as the basis of molecular biology.

for progress in the field of mass spectrometry-based proteomics (Listgarten and Emili, 2005).

There are numerous steps in a computational proteomics workflow that require fast and accurate algorithms. In Chap. 2, we give a summary of the most important computational problems in such a workflow. The remainder of this chapter is structured as follows: we will give a brief introduction to protein chemistry and explain some fundamental chemical terms that will be used in the following chapters. The last section of this chapter gives an outline of this thesis.

1.2 The Basics of Molecular Biology

The main focus of this text are computational methods and their application to mass spectrometry data. However, we need to introduce some key biological terms to make the next chapters understandable. We will not spend too much time explaining chemical or biological details, but try to give the reader enough background knowledge to put the remainder of this text into context.

1.2.1 The Central Dogma of Molecular Biology

One of the foundations of modern molecular biology is the *central dogma of molecular biology*, illustrated in Fig. 1.1. The central dogma was first enunciated by Francis Crick (Crick, 1970). In short, it states that the transfer of information in the cell goes from the DNA sequence over RNA to proteins and is unidirectional. The DNA is a polymer molecule which exists in all known organisms, including some viruses. Its main purpose is the storage of hereditary information. Subsets of the DNA, the genes, are transcribed into messenger RNA (mRNA), which is in turn translated into proteins. A conclusion from the central dogma was that proteins are the only molecules in the cell with a regulatory function, whereas RNA and DNA were assumed to be only carriers of information.

The biosynthesis of mRNA from DNA is called *transcription* and *translation* is the production of proteins from mRNA. At each step, there are chemical reactions and regulatory events that influence the function of the resulting protein. First, transcription and translation are imperfect processes and errors might occur. These errors might lead to mutant proteins, which might be dysfunctional or behave entirely different. Furthermore, the pre-messenger RNA undergoes a process called splicing, in which the original molecule is cut and reassembled. As illustrated in Fig. 1.1, *introns* are parts of the gene sequence that are removed during splicing and *exons* are parts that are assembled into the messenger RNA. Splicing can lead to several distinct transcripts or splice variants. In addition, proteins can receive several chemical modifications after translation. A protein can also excise different functions depending on its localization in the cell. Finally, proteins usually do not act by themselves but interact with proteins and other molecules to achieve their biological function.

In sum, there are numerous regulatory events that influence the information flow from DNA to proteins, and the complexity tends to increase the further we go downstream. Recent estimates state that there are roughly 20.000 human genes. Most of these genes are spliced: there are on average 4-5 splice variants of human genes (Le Texier *et al.*, 2006) (some genes have significantly more). Most mRNA molecules are further chemically modified, e.g., by adding a poly(A) tail or the 5' Cap. Many of these regulatory events can be analyzed using microarrays (Lee and Roy, 2004) or deep-sequencing (Sultan *et al.*, 2008). But these methods cannot monitor regulatory events that occur post-translational, e.g., after the biosynthesis of the protein. Right now, mass spectrometry is the method of choice to analyze these events, such as chemical protein modifications, protein localization or protein interactions (Rappsilber and Mann, 2002).

Finally, the central dogma itself has been challenged for several years (Gilbert, 1986). In fact, it is now widely accepted that there are many more classes of RNA molecules than previously known. It also became evident that RNA molecules are actually not only carriers of information but also have an independent and highly complex biological function. Nevertheless, it is reasonable to assume that proteins are still crucial to obtain a better understanding of the living cell.

1.2.2 Protein Chemistry

A protein is a chain of amino acids and its behavior and function are largely determined by this amino acid sequence. Peptides are shorter proteins, usually derived by enzymatic digestion of a protein. An amino acid is a molecule with both amine and carboxyl functional groups, as illustrated in Fig. 1.2. There are 20 standard or proteinogenic amino acids. These are the amino acids that are found in proteins and that are coded for in the genetic code. There are also several amino acids that are not proteinogenic such as hydroxyproline or carnitine.

Amino acids have the general sum formula $H_2NCHRCOOH$ where H_2N is the amine group, $COOH$ the carboxyl group and R is an organic substituent (side chain). The amine and carboxyl groups are bound to the same carbon, the so-called α-carbon atom. The individual amino acids differ in the substituent attached to this atom. Furthermore, they can be divided into several groups, depending on their biochemical properties such as hydrophobicity or acidity.

Amino acids are joined to peptides and proteins by a chemical reaction which results in a peptide-bond, i.e. an covalent bond between the amine group of one amino acid and the carboxyl

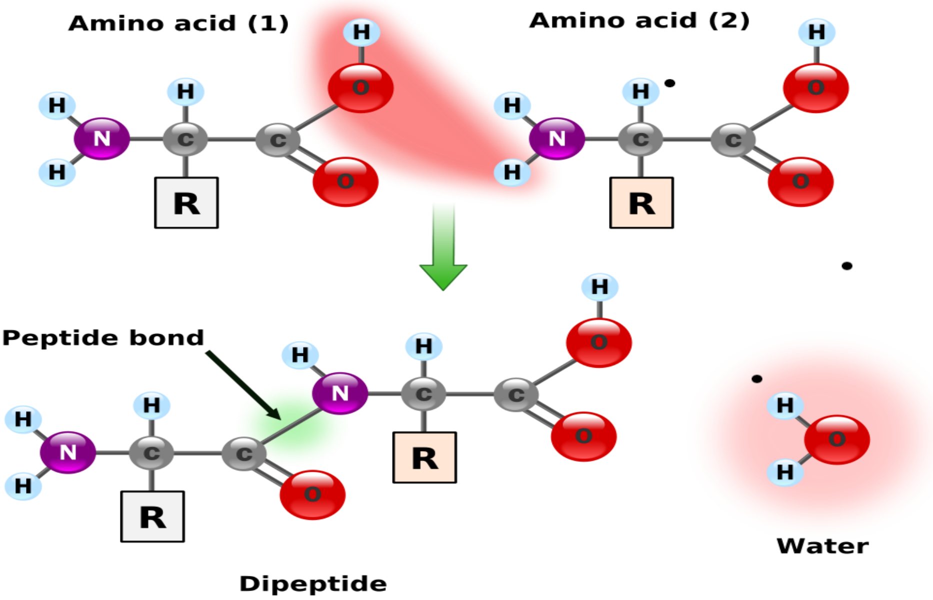


Figure 1.2: Proteins and peptides are sequences of amino acids. The amino acids are joined through a peptide bond. The illustration shows the formation of a peptide bond between two amino acids. It is a dehydration reaction and it results in the loss of a molecule of water. The illustration is taken from the corresponding wikipedia article at http://en.wikipedia.org/wiki/Peptide_bond, accessed August 2008.

group of the other amino acid. This reaction includes the removal of a water molecule from the growing amino acid chain. It is therefore a *dehydration* (or condensation) reaction. It can occur repeatedly and result in proteins which contain thousands of amino acids.

This chemical reaction is part of the translation of mRNA into proteins. The translation occurs in molecules called *ribosomes*. These molecules are situated in the cytoplasm of prokaryotic and eukaryotic cells. In short, the mRNA is decoded into an amino acid sequence according to specific rules, defined by the *genetic code*. The mRNA is a sequence of ribonucleotides and each triplet of ribonucleotides (codon) encodes for an amino acid. The transfer RNA (tRNA) is another, non-coding, RNA molecule that is involved in translation. As its name suggests, it transports amino acid molecules to the ribosome. Translation stops if a stop codon (UAA, UAG, or UGA) in the mRNA is encountered and the newly synthesized protein leaves the ribosome. In the aqueous environment of the cell, the protein folds in complex ways and any further chemical modification will influence its folding. The end of the amino acid chain with a free carboxyl group (-COOH) is called the C terminus, whereas the end with the free amine group (-NH2) is called the N terminus. This is also the end which is synthesized first during translation. In fact, translation is a more complex and energy-intensive reaction, but we will not give further details here and refer the interested reader to standard biology textbooks.

Apart from its amino acid sequence, the function of a protein is also strongly determined by *post-translational modifications* (PTMs). A PTM is the chemical modification of a protein after its translation. In many cases, a protein changes its function after modification or is even inactive without it. A PTM is usually a functional group that is attached to one of the amino acids in the protein, but addition of amino acids to the ends of the protein or a cut of the protein in the middle also counts as PTM. So far, the exact number of possible modifications is unknown, but the DeltaMass database (http://www.abrf.org/index.cfm/dm.home) lists 253 post-translational modifications (August 2008). This number is a conservative lower bound. Furthermore, many proteins have several modification sites, which adds a further layer of complexity.

The best-known PTM is probably phosphorylation, the addition of a phosphate (PO_4) group to a protein. The phosphorylation of proteins is a crucial regulatory mechanism. It occurs in both prokaryotic and eukaryotic organisms and many enzymes are activated only after phosphorylation. The phosphorylation usually occurs on serine, threonine, and tyrosine amino acid residues. The addition of the phosphate group can turn a hydrophobic portion of the protein into a polar and hydrophilic one. In this way, phosphorylation can introduce a change in the structure of the protein. We call the enzymes involved in this reaction *kinases* (which induces phosphorylation) and *phosphatases* (induces dephosphorylation).

Many proteins belong to the functional class of *enzymes.* These are molecules that catalyze chemical reactions and increase their reaction rate. Almost all reactions in the cell require enzymes to occur. Since enzymes are extremely selective and speed up only few reactions, the set of enzymes contained in a cell determines which reactions will occur in that cell.

Proteins can also have structural or mechanical functions, such as actin and myosin in muscles and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signaling, immune responses or cell adhesion (the binding of a cell to other cells or surfaces).

Another term occurring frequently in this thesis is *enzymatic digestion.* This is a chemical reaction catalyzed by enzymes such as trypsin or pepsin, in which proteins are cut into peptides. These cuts usually occur at the peptide bonds and most enzymes cut preferentially after certain amino acids. Note that in a mass spectrometry experiment, we usually deal with peptides, since the masses of most proteins are larger than the detection limit of a typical mass spectrometer.

Finally, we need to introduce a general term: an *ion* is a molecule which has either lost an electron or received an addition electron and is thus either positively or negatively charged. This is important, as mass spectrometry can only analyze charged molecules. Sometimes, the same peptide or protein will occur in different charge states. We will also speak of the *charge variants* of a peptide, if we want to emphasize this fact.

1.3 Thesis Outline

The remainder of this thesis is structured as follows: the second chapter introduces key concepts of mass spectrometry instrumentation and highlights computational challenges in the analysis of mass spectrometry data. We also introduce OpenMS, the software framework used to develop the software tools described in this thesis.

In the third chapter, we describe a novel algorithm to detect and quantify the signals of peptides in mass spectra. Our method is based on the combination of a wavelet function which approximates isotope peak intensities, the sweep-line paradigm from computational geometry which allows us to efficiently trace the monoisotopic m/z across several spectra, and a two-dimensional averagine template which combines a model of isotope peak pattern and an elution profile. A summary of this chapter was published in Schulz-Trieglaff *et al.* (2007).

The fourth chapter presents different applications of our method. We apply it to quantify myoglobin from human blood serum and to detect and quantify peptides in complex mixtures. Furthermore, we introduce LC-MSSim, a simulation software for mass spectrometry data used to generate benchmark data sets and to perform a detailed comparison of our approach to existing algorithms. This chapter contains material that was previously published in Schulz-Trieglaff *et al.* (2008a,b)

The fifth chapter deals with the quality assessment of mass spectrometry data. We present a framework for the quality control and assessment of large scale mass spectrometry experiments and evaluate it on real and simulated data. At the time this thesis is written, this chapter is submitted for publication (Schulz-Trieglaff *et al.*, 2008c).

In Chap. 6, we conclude this thesis with a summary of our work and an outlook on future directions.

Chapter 2

Mass Spectrometry-based Proteomics

Synopsis: *We give an introduction to mass spectrometry and its applications in research for modern biology and, more specifically, proteomics. Furthermore, we highlight the key computational problems in the computational analysis of mass spectrometry data.*

2.1 Introduction

The foundations of mass spectrometry (MS) have been developed as early as 1899 by the Germany physicist Wilhelm Wien. He devised an instrument that could separate positively charged ions by their charge-to-mass ratio (z/m) using electric fields (de Hoffmann and Stroobant, 2004). This method was further refined by the English scientist Sir Joseph John Thompson in the early 20th century. He also discovered the fact that most chemical elements exist in different isotope states using a variant of the instrument devised by Wien.

Compared to those early beginnings, the application of mass spectrometry to complex proteomic samples and in a high-throughput fashion have just recently become popular in the scientific community. To give an example, in 2002, John Fenn and Koichi Tanaka received the Nobel Prize in Chemistry for the development of Electrospray Ionization (ESI) and of Soft Laser Desorption (SLD), respectively. SLD was later improved by the German scientists Franz Hillenkamp and Michael Karras and is now more widely known as matrix-assisted laser desorption/ionization (MALDI). Both methods have greatly facilitated the analysis of proteins and peptides and will be explained in more detail in the later course of this chapter.

A typical mass spectrometry experiment in a modern laboratory involves the analysis of mixtures i.e. samples containing many proteins at different concentrations. For the remainder of this thesis, we have to keep in mind that the majority of the low-abundance proteins are not observed in these experiments. High-throughput mass spectrometry experiments capture only a subset of all proteins contained in the sample. Furthermore, there is even some stochastic moment to these experiments as reproducibility between technical replicates can be low and not all proteins will be detected every time.

This thesis focuses on data generated from liquid chromatography-mass spectrometry (LC-MS) experiments. This is a special type of mass spectrometry experiments in which the complexity of a sample is reduced by coupling a chromatographic column to the mass spectrometer. The column separates the sample compounds by some chemical property and thus not all compounds enter the mass spectrometer at the same time. Figure 2.1 illustrates a generic LC-MS experiment.

The outcome of an LC-MS experiment is an *LC-MS map*, a set of mass spectra. Each spectrum is a set of two-dimensional data points $p_i = (m/z, it)$. m/z is the mass-to-charge ratio and has the unit thomson (Th). *it* is the intensity. Its exact meaning depends on the instrument used, but in most cases we assume it to be an *ion count.* We use it as an estimate of relative abundance of the compound at the corresponding m/z. Note that the signal intensities of the detected peptides are only comparable between mass spectrometers of the same type and if care is taken during sample preparation.

In this chapter, we introduce the components of the LC-MS setup and provide a brief overview of the most important technologies used in each step. This is important as different instruments give us data with different characteristics. On this account, we will also put more emphasis on the influence that each technology has on the mass spectra it generates, than on details of each technology.

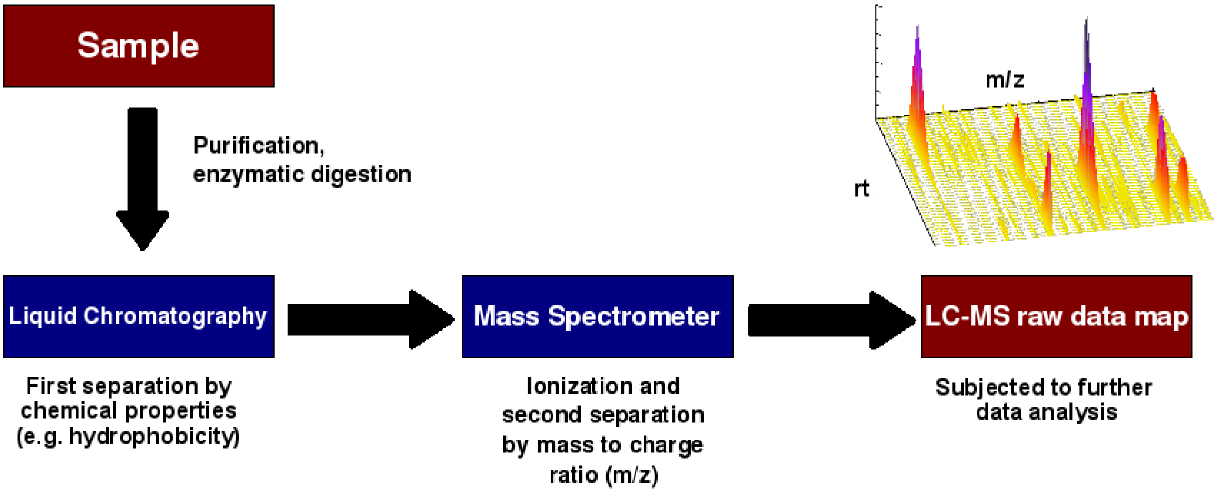

Sample
Purification, enzymatic digestion
Liquid Chromatography
First separation by chemical properties (e.g. hydrophobicity)
Mass Spectrometer
Ionization and second separation by mass to charge ratio (m/z)
LC-MS raw data map
Subjected to further data analysis
m/z
rt

Figure 2.1: *A generic LC-MS experiment. It consists of sample preparation, chromatographic separation and mass spectrometry analysis. The resulting set of spectra is the raw LC-MS map.*

2.2 Principles and Instrumentation

2.2.1 Sample Preparation

In general, we assume that our sample is in a liquid state. Solid samples need to be disrupted and solved in a solvent. The sample consisting of proteins and other substances is subjected to enzymatic digestion. This facilitates the later analysis as MS of whole proteins is less sensitive than peptide MS (MacCoss and Matthews, 2005). This is also called the *bottom-up approach*. In most cases, the enzyme *trypsin* is used for digestion since it cuts the amino acid sequence of a protein at well defined positions. The generally accepted *Keil rule* states that trypsin cleaves next to arginine or lysine, but not before proline, although this has been challenged recently (Rodriguez *et al.*, 2008). So-called *miscleavages* might also occur, leading to peptides than span one or two tryptic cleavage sites.

Other enzymes are used to remove biomaterials such as lipids or sugars. Another step in a typical sample preparation is to break protein-protein interactions using urea or to reduce disulfid bonds between cysteine residues. Finally, we need to be aware that we usually deal with mass spectral signals of peptides and not of full proteins. Thus, to infer the abundance or sequence of the whole protein requires an additional computational step.

2.2.2 Chromatographic Separation

After enzymatic digestion, we have obtained a mixture of peptides and maybe other sample compounds. If we would inject this sample directly into a mass spectrometer, we would obtain just one crowded mass spectrum containing many, and often overlapping, signals of all sample compounds. To alleviate this effect, we perform a simplification of the sample using liquid chromatography as a first step.

The basic principle of chromatography is simple: the sample molecules traverse the length of the chromatographic column. They are retarded by chemical or physical interactions with the column material, the stationary phase. The amount of retardation depends on the nature of the molecules, the stationary phase and the solvent employed. The time at which a compound elutes is called the *retention time*, abbreviated rt. For peptides, the retention time is heavily influenced by their amino acid sequence. Peptides with similar amino acid compositions will thus elute at similar retention times. It is also possible to combine several steps of chromatographic separation. Its application to proteomics was pioneered by the Yates lab at the Scripps Research Institute in San Diego, La Jolla, US (Washburn *et al.*, 2001; Wolters *et al.*, 2001). If applied to protein samples, it is called *Multidimensional Protein Identification Technology* or MudPIT.

To sum up, the peptides and all other sample components are separated by liquid chromatography and only after this separation, they are injected into the mass spectrometer. This obviously leads to less complex subsets of the sample and to less crowded mass spectra. The drawback is that the data size increases and that we will obtain many mass spectra per sample, often several thousand instead of just one. Each spectrum represents a subset of the sample compounds as the elute from the column. As stated above, the set of all spectra obtained from a sample using one combination of column and mass spectrometer is the *LC-MS map*. Figure 2.2(a) gives a bird's eye view of a map.

2.2.3 Mass Spectrometry Instrumentation

After the chromatographic separation, the sample molecules are injected into the mass spectrometer. It consists of an *ion source*, a *mass analyzer* and a *detector*. Electrospray Ionization (ESI) and Matrix-Assisted Laser Desorption/Ionization (MALDI) are the two most commonly employed ion sources for proteins and peptides.

Ion Sources

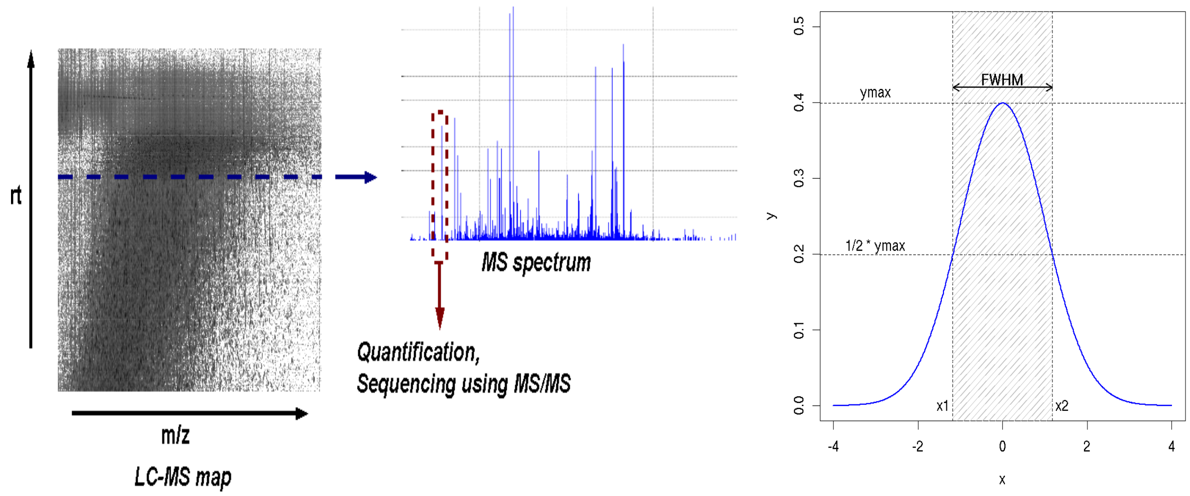
ESI ionizes the analytes out of an aqueous solution and is therefore the ideal method for LC-MS setups. The liquid containing the analytes is pushed through a small capillary to which a potential difference is applied (Fenn *et al.*, 1989). This causes a strong electric field which in turn causes an accumulation of charged molecules at the surface of the liquid. Since equal charges repel, the liquid pushes itself out of the capillary. It forms an aerosol, a mist of of small liquid droplets. The solvent molecules, which are usually more volatile than the analyte, evaporate and force the analyte molecules into closer vicinity. The molecules repel each other and break up the droplets. The process repeats until all solvent molecules are removed and the analytes form lone ions. Interestingly, even though this method was invented 19 years ago, and its underlying principles are even older, there is still an ongoing debate on the exact mechanisms of this ionization process (de Hoffmann and Stroobant, 2004).

ESI is well suited for complex samples and high-throughput experiments. Consequently, this thesis focuses on data from ESI ion sources. From a data analysis point of view, it is important to realize that using ESI, copies of the same peptide can receive different charge states, i.e. the same peptide can appear with charge 2, 3 and 4, just to give an example. This complicates the data analysis.

An ESI source can be operated in positive or negative ion mode. In positive ion mode, the compounds receive a proton as charged adduct, in negative mode they receive an electron. For tryptic peptides, most laboratories perform ESI mass spectrometry in positive ion mode, since the tryptic digest leads to peptides with negatively charged terminal amino acids. More generally, the mass-to-charge ratio (m/z) of a sample compound in ESI mass spectrometry is given by

$$\frac{m}{z} = \frac{M + im_a}{ie} \tag{2.1}$$

where M is the parent mass (the mass of the peptide), i is the number of charges and e is the elementary charge. m_a is the mass of the charged adduct. In most cases, this adduct is either



(a) Bird's eye view of an LC-MS map

(b) The FWHM of a peak

Figure 2.2: *Left: a bird's eye view of an LC-MS map and a single mass spectrum. Each spectrum has an associated retention time and represents a subset of the sample. Right: illustration of the Full-Width-At-Half-Maximum (FWHM) of a peak. It is defined as the difference of the m/z values at which the peak intensity is at the half of its maximum.*

an electron or a proton, but other adducts such as sodium (Na^+) or potassium K^+ are possible.

From a data analysis point of view, it is also important to realize that the ionization process is not perfect. Due to *competitive ionization*, some molecules are easily ionized whereas others do not ionize at all or cannot even be solved in liquid. These and other effects will cause the loss of some molecules before they even reach the detector. We need to keep in mind that using LC-MS we will only observe a subset of all sample compounds.

For reasons of completeness, we mention MALDI as well which is a method to ionize molecules out of a dry, crystalline matrix via laser pulses. It is widely used to identify peptides in mixtures and but is not well-suited for quantitative measurements (Mann and Aebersold, 2003) and thus we did not use any MALDI spectra for this thesis.

The Mass Analyzer

The mass analyzer measures the mass-to-charge ratio (m/z) of the ionized analyte. For our purposes, its key parameters are sensitivity, mass resolution and mass accuracy (Mann and Aebersold, 2003). The sensitivity characterizes the ability of the mass analyzer to detect weak signals. Mass resolution and mass accuracy describe how well the analyzer is able to resolve peaks with similar mass and how accurately it measures this mass, respectively. The mass resolution is a dimensionless unit and is expressed as the ratio of the mass of a signal peak in a mass spectrum and its Full-Width-At-Half-Maximum (FWHM). The FWHM of a peak is illustrated in Fig. 2.2(b).

Unfortunately, this is not a formal definition, but only a convention. The mass resolution is sometimes also calculated using the peak width at certain intensity thresholds or by measuring the intensity in a valley between two peaks. We will restrict ourselves to the first definition for the remainder of this text.

Finally, the mass accuracy is defined as the observed difference between the real mass of an analyte and the mass measured in the spectrum :

$$\text{mass accuracy} = \text{mass}_{\text{real}} - \text{mass}_{\text{measured}}. \tag{2.2}$$

The mass accuracy is often expressed in parts per million (ppm) :

$$\text{ppm} = 10^6 \times \text{mass accuracy}/\text{mass}_{\text{measured}}. \tag{2.3}$$

To give an example, if we know that the true mass of a compound is 1000.0 m/z and our mass spectrometer measures a signal for this compound at 999.99 m/z, then the accuracy of this measurement was $\approx$ 10 ppm. Zubarev and Mann (2007) suggested that this definition of mass accuracy is actually a mass deviation but the use of the term accuracy is wide spread.

There are four basic types of mass analyzer currently used in proteomics research. These are the Time-Of-Flight (TOF), Quadrupole, Fourier Transform Ion Cyclotron (FT-ICR) and the Orbitrap mass analyzer, each with its own strengths and weaknesses.

The TOF analyzer relies on a simple physical principle. It uses an electric field to accelerate the ions and then measures the time t they take to reach the detector. The ions with charge z are accelerated by a potential V. In this case, it holds that

$$t^2 = \frac{m}{z} \cdot (\frac{d^2}{2eV}) \tag{2.4}$$

where d is the length of the flight chamber and e is the elementary charge (de Hoffmann and Stroobant, 2004). Thus, the mass-to-charge ratio m/z of an ion stands in a quadratic relationship to the flight time t and can readily be calculated.

This first mass spectrometers used TOF mass analyzers (Wiley and McLaren, 1955), and they are nowadays a well-established technology, reliable and relatively cheap. Nevertheless, new technologies have been invented in recent years and most of the data that we use in this thesis are not generated by TOF instruments. Therefore, we will briefly describe the basics of other mass analyzers as far as their are important for this work.

The Quadrupole analyzer is a mass filter: it is capable of transmitting only ions with a specified m/z. It consists of four parallel electrodes. Each pair of opposing electrodes is connected electrically and a radio frequency voltage is applied between both pairs of electrodes. A direct current voltage is superimposed on the radio frequency voltage. After leaving the ion source, ions fly through the Quadrupole in between the electrodes. Only ions with a certain m/z will have a stable trajectory and can pass through. This allows selection and trapping of ions with a particular m/z but also to spectrum over a m/z range by changing the voltages. As an example, in the Quadrupole Ion Trap or Paul Trap, the ions are trapped for a certain time interval. Its basics were firstly described by Paul and Steinwedel (1953). Depending on the radio-frequency voltage applied, ions of a specific m/z can be trapped in the space between the electrodes. After that, the strength of the electric field is changed such that the ions are catapulted out of the trap and have their m/z recorded. Ion traps have the reputation of being very robust and are inexpensive (March, 2000). They produced much of the proteomics data which is available today. Their disadvantage is the low mass accuracy: the trap can only capture a limited number of ions before they distort the accuracy of the mass measurement. Recent developments try to circumvent this problem by increasing size and changing the shape of the ion trap, such as in the two-dimensional or linear ion trap, abbreviated as LTQ ("Linear Trap Quadrupole"). The Quadrupole can also be combined with a Time-Of-Flight analyzer, this combination is called QTOF (sometimes also written qTOF). In this setting, the quadrupole serves as an ion guide for the TOF analyzer. This results in higher sensitivity and mass accuracy for the TOF instrument (Chernushevich *et al.*, 2001). The microTOF instrument series produced by Bruker Daltonics is an example for a QTOF instrument. We will use data from this instrument in the following chapters.

The Fourier Transform-Ion Cyclotron (FT-ICR) instrument is also a trapping mass spectrometer. FT-ICR was invented by Comisarow and Marshall (1974). It determines the mass-to-charge ratio of ions based on the cyclotron frequency of the ions in a fixed magnetic field. This field is created under high vacuum in a Pienning ion trap. Once they are in the trap, the ions are excited to a cyclotron radius by an oscillating electric field perpendicular to the magnetic field in the trap. The mass-to-charge ratio determines the cycling frequency of an ion which is measured by detectors placed at fixed positions in the trap. The strengths of an FT-ICR mass analyzer are its high sensitivity, mass accuracy (1-2 ppm) and resolution. Recent developments include the combination of a Fourier Transform mass spectrometer with a Linear Trap Quadrupole (Syka *et al.*, 2004a).

The LTQ Orbitrap is the most recently developed analyzer (Hu *et al.*, 2005). The Orbitrap operates by trapping ions around a central spindle electrode. The suggestive name derives from the ions traveling in an orbit around the central electrode in the trap. An outer barrel-like electrode is placed coaxially with the central electrode. The ions are tangentially injected into the trap and oscillate around the central electrode. In addition, they also move back and forth along the axis of the electrode. Their m/z values are calculated from the frequency of their oscillations by applying the Fourier Transform. Orbitraps have a mass accuracy and resolution comparable to FT-ICR instruments but at a fraction of the costs (Scigelova and Makarov, 2006). Consequently, more and more laboratories use them for their experiments.

The Mass Detector

Finally, the mass detector counts the number of ions at each m/z. Possible detectors are *photographic plates*, *faraday cylinders* or *array detectors* (de Hoffmann and Stroobant, 2004). Most detectors need some time to recover after an ion hit. This time span is called *dead time* (Chernushevich *et al.*, 2001). This means that ions with very similar m/z might not be accurately recorded.

2.3 Mass Spectrometry-based Proteomics

We introduced the LC-MS setup and the components of a mass spectrometer. The outcome of an LC-MS experiment is the LC-MS map. The map is a collection of spectra where each spectrum contains signals of sample compounds eluting from the chromatographic column at a specific time point. Mass spectra are complex measurements which contain a significant amount of noise.

This thesis deals with the field of proteomics, the study of cellular function at the protein level. For this reason, we are mainly interested in proteins and peptides and less in other molecules that might be contained in a biological sample. As outlined above, proteins are usually digested before analyzing them using mass spectrometry and we will from now on only speak of peptides and not of proteins. Fortunately, peptides exhibit well predictable signals in a mass spectrum. Chapter 3 introduces *isotopic peak distributions* and presents an algorithm to efficiently locate and quantify peptides in mass spectra.

Proteomics has raised great expectations for the discovery and quantification of protein biomarkers for improved diagnosis or stratification of a wide range of diseases, including cancer (Mann and Aebersold, 2003). Blood plasma and other body fluids are expected to be excellent sources of these biomarkers because they circulate through, or come in contact with, a variety of tissues – with all tissues in the case of plasma (Zhang *et al.*, 2005). But all these applications require an accurate detection of peptide signals and their accurate comparison across different experiments. Due to the large amount of data generated in an LC-MS run, there is a pressing need for efficient algorithms in this field. In the following sections, we will give an overview of the key steps in a LC-MS data analysis workflow.

From the data generated in an LC-MS experiment, we can determine the mass, charge and retention time of the peptides contained in the sample under examination. But for most applications, this is not sufficient and we are also interested in the actual amino acid sequences. This requires an additional measurement, the acquisition of MS/MS or tandem MS spectra. We will briefly explain the principle of peptide sequencing using tandem mass spectrometry since we will refer to this technology from time to time in the later chapters of this thesis.

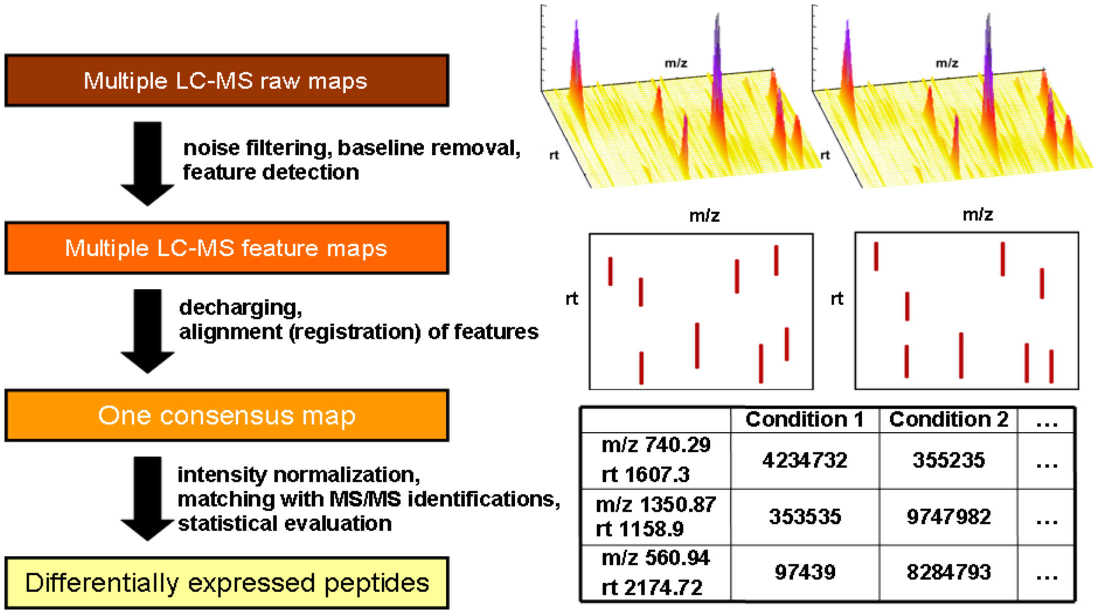
An MS/MS spectrum provides additional information for selected signals in the LC-MS map. The MS/MS spectrum is a mass spectrum and contains m/z and intensity measurements of fragments of the precursor ion. The mass spectrometer usually records MS/MS spectra automatically for signals with an intensity higher than a given threshold. This procedure is called *Data-dependant Acquisition* (DDA) and the compounds selected for MS/MS recording are called *precursor ions.*

There are different approaches to perform this fragmentation. The most frequently used method is *Collision Induced Dissociation* (CID) (Mitchell Wells and McLuckey, 2005), also called *Collisionally Activated Dissociation* (CAD). In short, the mass spectrometer captures selected ions, the *parent* or *precursor ions*, accelerates them using an electric potential and allows them to collide with neutral gas molecules, such as helium or argon. This results in a breakage of the parent ion into smaller fragments. These fragments are recorded in a MS/MS spectrum. Recently, new fragmentation methods have emerged such as *Electron Transfer Dissociation* (ETD) or *Electron Capture Dissociation* (ECD) (Syka *et al.*, 2004b; Zubarev *et al.*, 1998).

If peptides are subjected to CID, they usually break at the peptide bonds. The mass spectrum records the m/z values of the fragment masses in the MS/MS spectrum. The mass differences between consecutive fragments match the corresponding amino acids. In many cases, this allows at least a partial reconstruction of the amino acid sequence (*de-novo sequencing*). If this is not possible due to noise, missing fragments or unexpected post-translational modifications, we can compare the tandem spectrum against a library of manually annotated spectra(*spectral library searching*) (Lam *et al.*, 2007). Another possibility is to compute theoretical MS/MS spectra for peptides obtained from a protein sequence database and to compare them against the experimental spectrum (*sequence database searching*). The latter method is the most common one. Popular algorithms are, for example, Sequest (Yates *et al.*, 1995), Mascot (Perkins *et al.*, 1999) or InsPecT (Tanner *et al.*, 2005).

This is of course only a rough sketch of a complex process. But for the remainder of this thesis, it suffices to know that we can, in theory, obtain the amino acid sequences for the peptides in our sample. This is a difficult task, since peptide identification using MS/MS is error-prone and the algorithms available today are known to produce high false positive rates. We have to be aware that some of the obtained amino acid sequences will be wrong, others might only partially be correct.

It is also important to realize that most MS/MS data available today was acquired in an automatic fashion, using the abovementioned Data-dependent Acquisition. In this case, the


Multiple LC-MS raw maps
noise filtering, baseline removal, feature detection
Multiple LC-MS feature maps
decharging, alignment (registration) of features
One consensus map
intensity normalization, matching with MS/MS identifications, statistical evaluation
Differentially expressed peptides
m/z
rt
m/z
rt
m/z
rt
m/z
rt

	Condition 1	Condition 2	...
m/z 740.29 rt 1607.3	4234732	355235	...
m/z 1350.87 rt 1158.9	353535	9747982	...
m/z 560.94 rt 2174.72	97439	8284793	...

Figure 2.3: *A typical LC-MS data analysis workflow. Key steps comprise peptide feature extraction, LC-MS map alignment and statistical analysis. The result is a table with expression measurements for each peptide in each condition.*

mass spectrometer selects prominent signals in each spectrum automatically for fragmentation. This usually leads to an under-sampling of peptide ions (Liu *et al.*, 2004) since weak signals will be missed. It is almost impossible to sequence every peptide in a sample in a single-pass analysis.

2.4 Mass Spectrometry Data Analysis

In the previous sections, we explained the setup and the different components of Liquid Chromatography coupled to Mass Spectrometry. The next chapter will address a specific task in mass spectrometry-based proteomics, the quantification of peptides from LC-MS data. But this is only one among several steps in a LC-MS data analysis workflow. In this section, we will briefly highlight the major problems in a typical workflow and discuss how our work fits into a larger pipeline (see Fig. 2.3). A good overview is given in Listgarten and Emili (2005). At this point, we will only present the most important computational steps.

2.4.1 Low-level Signal Processing and Peptide Feature Discovery

As indicated in Fig. 2.3, the data analysis starts with the unprocessed (raw) LC-MS map. Since the map is subject to background chemical and electronic noise, together with systemic contaminants in the mobile phase in the chromatographic column, methods for noise reduction and signal enhancement are commonly applied. Typical low-level processing steps include smoothing using median, moving average filters or the Savitzky-Golay filter (Wang *et al.*, 2003).

In some cases, the LC-MS map will also contain noise with lower frequency, usually referred to as *baseline noise*. Methods that are applied to estimate and then subtract the baseline from a mass spectrum are iterative local regression (Wagner *et al.*, 2003), matched filter (Sauve and Speed, 2004) and Fourier transformations (Baggerly *et al.*, 2003).

The next step involves the detection and quantification of peptide signals in the LC-MS map. The computational challenge is to deal with data from different instruments, to recover weak peptide signals and to assign abundance and charge state correctly. Chapter 3 reviews previous work in this field and presents our approach, therefore we will not give more details here.

2.4.2 LC-MS Map Alignment

After the step of peptide feature detection, we know mass, charge and retention time of the peptides in the LC-MS map (see step 2 in Fig 2.3). We estimate their relative abundance by integrating their signal intensities. But the comparison of these abundances across different samples is not straightforward. It is hampered by the fact that the retention time at which a peptide elutes from the chromatographic column is not stable across different experiments. The elution behavior is distorted in complex ways by differences in column performance due to changes in ambient pressure and temperature.

Several algorithms were developed to correct these distortions. Examples are hidden Markov models (Listgarten *et al.*, 2007), dynamic time warping (Wang *et al.*, 2003), or clustering (Sauve and Speed, 2004). These algorithms usually make different assumptions about the type of distortions (e.g. linear or non-linear). Some of them can be applied to the raw spectra directly and others operate on the m/z and rt coordinates of peptide features.

If not stated otherwise, we will use an algorithm based on Geometric Hashing (Lange *et al.*, 2007). This algorithm is based on methods for shape and object recognition that stem from Computer Vision. It corrects for distortions in retention time between LC-MS maps by hashing the difference vectors between pairs of peptide signals in two maps. The most common difference vector yields a first estimate of the dewarping function. This transformation is further refined using piecewise regression between pairs of peptide coordinates.

The *decharging* of the feature coordinates is a step that can be performed after the alignment of retention times. The peptide features with mass-to-charge ratio m/z are simply mapped to the corresponding parent peptide mass and have their intensity measurements summarized. Obviously, this requires a correct charge estimate. The process of decharging can be a non-trivial step for complex samples.

2.4.3 Statistical Analysis

The actual detection of significant differences in peptide abundances is the next step. It involves the comparison of peptide signal intensities between the aligned LC-MS maps. The resulting statistical problems are comparable to issues in microarray data analysis and similar methods, such a linear models (Oberg *et al.*, 2008) or lowess smoothing for normalization (Sauve and Speed, 2004), are applied. Further computational steps include the mapping of peptide sequences obtained from MS/MS to features as well as the abundance estimation on the protein level.

2.5 OpenMS - Software Library for Computational Mass Spectrometry

In this thesis, we made heavy use of OpenMS, which is a software library that offers an infrastructure to develop algorithms for mass spectrometry data analysis (Sturm *et al.*, 2008). The library contains data structures to store and algorithms to process mass spectra. It includes a visualization tool for LC-MS maps that we used throughout this thesis. OpenMS is also the basis of a set of programs called TOPP, The OpenMS Proteomics Pipeline. Using TOPP, users can perform repeatedly occurring task in LC-MS data analysis such as visualization, file conversion and signal processing (Kohlbacher *et al.*, 2007). OpenMS is a joint project of research groups at the Free University of Berlin, Tübingen University and the University of Saarbrücken and it is available for free under the Lesser GNU Public License at www.openms.de.

The methods described in the following chapters are part of OpenMS Release 1.0. Since then, they have been further refined and improved by various developers in the OpenMS project. The most recent OpenMS release version at the time this thesis was written is 1.2. We give details on the implementation of our algorithms in OpenMS in the appendix of this thesis. In addition, we used OpenMS in this thesis for various tasks such as file format conversion, filtering and visualization of spectra. We give implementation details and installation instructions for our tools in the Appendix.

2.6 Summary

We introduced the LC-MS setup and explained the different components of a mass spectrometer. These components are the ion source, the mass analyzer and the detector. The most commonly employed ion sources for peptides and proteins are Electrospray Ionization (ESI) and Matrix-Assisted Laser Desorption/Ionization (MALDI). The focus of this text are LC-MS data generated from ESI-MS instruments. Their key characteristic is that instances of the same peptide can appear as different charge variants (ions). The mass analyzer determines the mass-to-charge ratios of the ionized sample compounds whereas the detector counts the number of ions at each m/z.

Key parameters of a mass spectrometer are sensitivity, mass resolution and mass accuracy (Mann and Aebersold, 2003). The sensitivity is the ability of the mass analyzer to accurately detect weak signals. Mass resolution and mass accuracy describe its ability to resolve signals with similar mass and how accurately it measures this mass, respectively. The mass resolution is a dimensionless unit, expressed as the ratio of the mass of a signal peak in a mass spectrum and its Full-Width-At-Half-Maximum (FWHM).

In addition, we presented some challenges in the computational analysis of LC-MS data sets. A typical workflow consists of low-level processing of the mass spectra including smoothing, baseline reduction and feature discovery. Subsequent steps are alignment of the feature sets, normalization, difference detection, mapping of MS/MS-derived sequences to features and mapping of peptide sequences and abundances to the protein level. In short, every step in a typical data analysis workflow requires tailored algorithms and statistical methods. This thesis focuses on the low-level processing of the LC-MS data, such as feature discovery and quality assessment of the raw LC-MS maps.

The next chapter presents a new algorithm for the detection and quantification of peptide signals from mass spectrometry data. We review previous work on this topic and compare our approach to existing ones.

Chapter 3

An Algorithm for the Quantification of Peptides from LC-MS Data

Synopsis: *This chapter presents a novel method for the quantification of peptides from LC-MS data. We introduce the problem of peptide quantification and review previous work. Our approach improves on existing algorithms in several aspects: it applies a filtering step based on a tailored mother wavelet, includes a refinement step using a two-dimensional model of a peptide signal and makes efficient use of the redundancies inherent in LC-MS runs to improve the signal detection process. We conclude the chapter with three computational experiments demonstrating the abilities of our algorithm.*

3.1 Introduction

This chapter deals with an important preprocessing step in a quantitative mass spectrometry experiment: the computational detection and quantification of the peptide signals in an LC-MS map.

As we outlined in the previous chapter, mass spectrometry-based proteomics is a relatively new scientific field that has caught a lot of interest in recent years. The computational quantification of peptides and their modifications from LC-MS data has just started to emerge as a research field for the Bioinformatics community. Consequently, many data sets that are currently available are still of moderate quality. Instrument resolution and mass accuracy are often poor since low resolution mass spectrometer are still the working horses in many laboratories. But this severely hampers a computational analysis. On the other hand, the capabilities of analytical instruments are evolving at a rapid pace, and high mass resolution instruments are becoming more and more common. It is therefore important to develop algorithms that are independent of data quality but also instrument type employed.

The advantage of our approach is that it is flexible and can be applied to data from different instruments and of different mass resolutions. It consists of two main parts: a seed detection step that uses a tailored mother wavelet to rapidly detect potential isotope pattern and a filtering step using a two-dimensional model of a peptide signal.

The structure of this chapter is as follows: we introduce the problem of peptide quantification in LC-MS maps and give an overview of previous work in this field. Furthermore, we explain some mathematical and chemical terms that are necessary to put our work into context. Finally, we describe our algorithm for peptide signal detection and quantification. This description has in parts been already published (Schulz-Trieglaff *et al.*, 2007, 2008a). Details and improvements of the wavelet function our algorithm uses are published in Hussong *et al.* (2007).

3.1.1 Peptide Signals in Mass Spectra

The focus of this chapter is on the detection of peptide signals in LC-MS maps. Peptide signals exhibit a characteristic shape in a mass spectrum. This helps us to distinguish them from noise or the signals of other sample compounds which are of lesser interest to us. The atoms contained in a peptide exist in different isotope variants which occur in a fairly constant distribution in nature. This gives rise to a characteristic pattern of adjacent peaks in the mass spectrum. We will refer to this peak intensity pattern as an *isotope distribution.* In Sect. 3.2.1 we explain how to compute this distribution for a given sum formula. Important for our purposes is the fact that we can approximate the isotope distribution of a peptide if we only know its mass and not the exact composition. In this case, we estimate this distribution using so-called *averagines* (Senko *et al.*, 1995a). The averagine represents an “average amino acid”. Its atomic composition is estimated by averaging the composition of many protein sequences obtained from public databases. We explain this concept in more detail in Sect. 3.2.2.

A peptide signal in an LC-MS map is not only characterized by its mass, charge and isotopic distribution but also by its behavior in the chromatographic column. It elutes over a certain interval of time from the column and can be observed in several consecutive spectra. Ideally, the elution profile should follow a normal distribution around its centroid, but fronting and tailing effects are frequently observed in practice (Di Marco and Bombi, 2001), so we have to consider asymmetric shapes as well.

These facts are best illustrated with an example. Figure 3.1 shows the signal of a peptide ion recorded with a high-resolution mass spectrometer. The left plot gives the two-dimensional

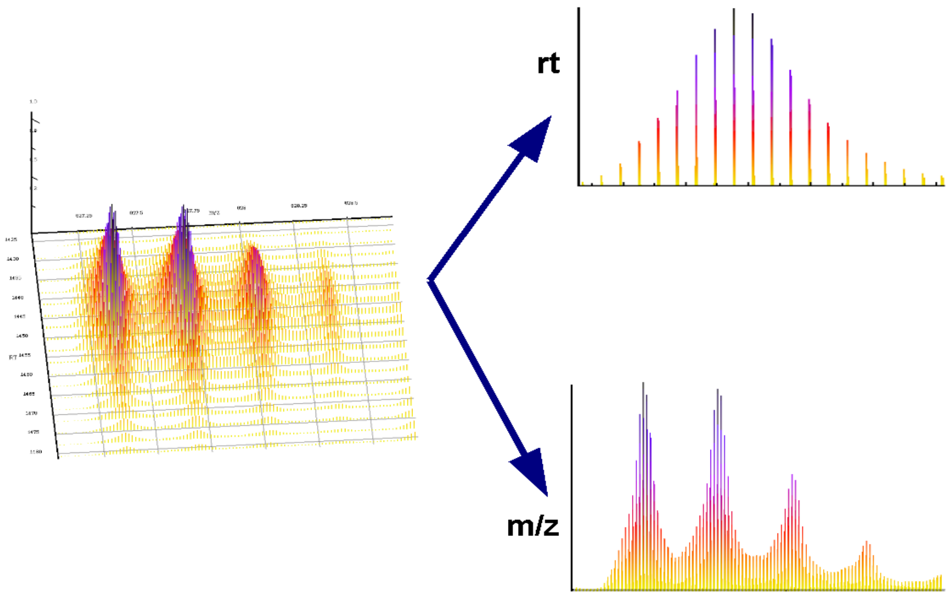


Figure 3.1: *The signal of a peptide in a mass spectrum. The left plot shows the two-dimensional signal of a peptide from a LC-MS map. The two plots on the right hand side give the projections of the signal on the m/z and retention time axis. The signal is caused by a peptide at m/z* 627.35*.*

signal, the two smaller insets show the projections on m/z and retention time axis, respectively. As we can see, the peptide has a monoisotopic m/z of 627.35. The centroid in retention time is approximately 1450 seconds.

We can also see that the signal consists of several peaks (local maxima). This is the above mentioned isotope pattern. For lower masses, the relative heights of the isotope peaks are approximately binomial-distributed, for higher masses this distribution will become more Gaussian-shaped. Note that we will make a difference between peaks and data points. A data point is a single intensity measurement in mass spectrum, characterized by m/z and intensity. A peak is always a local intensity maximum which might or might be not part of a peptide signal. We give more details on isotope peak patterns and their computations in Sect. 3.2.1.

If the mass spectra are recorded on a high-resolution mass spectrometer, the charge of the peptide ion can be determined from the m/z distances between consecutive isotope peaks. Since carbon is the most frequent element, the mass difference between its isotopes, 1.003355, dominates. Consequently, the distance between isotope peaks of an ion with charge z will be $\approx 1.003355/z$. The ion in Figure 3.1 has a charge of 3. The monoisotopic mass, that is the mass of the lightest isotope peak, of this peptide is thus $627.35 \times 3 = 1885.05$. We will give a more precise explanation of the monoisotopic mass in Sect. 3.2.1. If the mass resolution is low, or if the spectrum is crowded and noisy, charge determination is obviously more difficult. Section 3.2.3 gives an overview of the most important approaches.

To summarize, peptides exhibit a characteristic, two-dimensional pattern in an LC-MS map. Especially the m/z part of this pattern can easily be approximated. From now on, we will refer to the set of data points caused by one peptide ion as a *peptide feature*. Note that the term *feature* is very general and thus not an ideal choice. Nevertheless, it is well established in the proteomics community and if we use the term feature in the remainder of this thesis, we always mean a peptide feature.

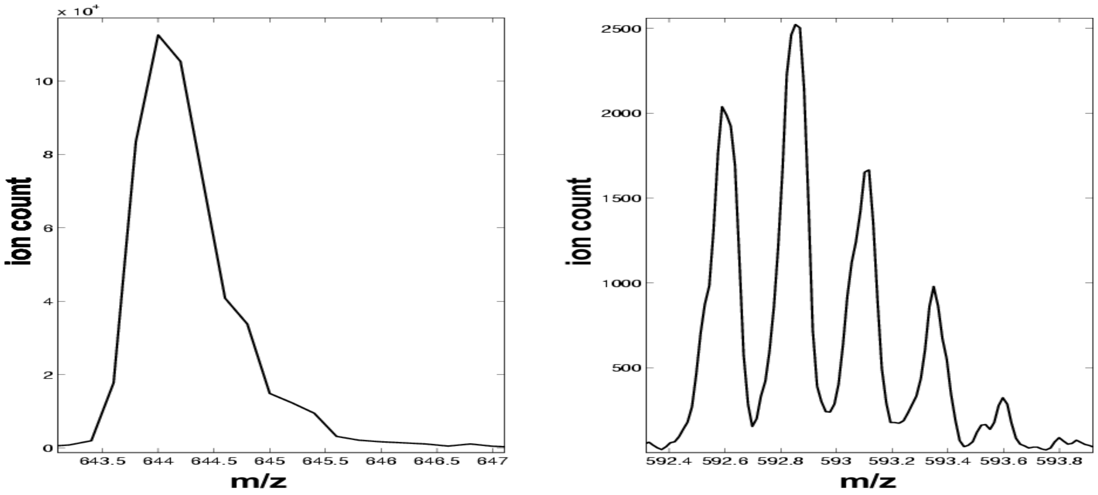


Figure 3.2: *A peptide signal measured at different mass resolutions. The signal to the left was measured on an older TOF Instrument, the right signal on a LTQ mass spectrometer.*

3.1.2 Peptide Quantification Problem

A central goal in functional proteomics is to globally detect changes in protein and peptide abundances in biological systems, providing a snapshot of the protein expression of a cell as it responds to biological perturbations such as diseases or drug treatment. To accurately estimate the abundances of all peptide constituents in an LC-MS map, we need to rapidly detect potential isotope patterns and, for each data point in their vicinity, decide whether it is caused by this particular peptide ion or not. The abundance of the peptide is subsequently estimated by summing the intensities of all data points caused by it and thereby approximating the intensity (ion count) of the signal.

The difficulty of the quantification problem lies in the wide range of mass resolutions of currently available mass spectrometers. Figure 3.2 illustrates this. The left signal is a peptide recorded on a Linear ion trap. The signal is a peptide ion, but determination of charge and monoisotopic mass is difficult since the isotope peaks are not resolved. The right signal was recorded an on a FT-Instrument. Here, isotope peaks are clearly separated but we need to account for other problems such as overlapping patterns. Low signal-to-noise ratios in the spectra and sample contaminants further hamper an accurate computational quantification.

The peptide quantification task can also be seen as a data reduction problem. An LC-MS map generated by an Orbitrap or Fourier Transform instrument can easily reach a file size of several gigabytes whereas the peptide lists are only a few kilobytes large. More complex analyzes such as the search for post-translational modification or mass labels can easily be performed on the reduced data set. From a computational point of view, these tasks reduce to a range query on the feature coordinates.

At the moment, there are two prevailing experimental techniques for a quantitative comparison of two or more samples based on the signal intensities in LC-MS. The first one relies on chemical or mass tag *labeling*. Peptides in two different samples are covalently modified by adducts which are chemically similar but have a known, observable difference in mass. The samples are subsequently proteolysed and mixed. Relative changes in abundances can be determined from the ratios of ion counts of the differentially labeled peptide pairs. Commonly used tags are isotope-coded affinity tag (ICAT) (Shiio and Aebersold, 2006) or stable isotope labeling with amino acids in cell culture (SILAC) (Ong *et al.*, 2002). Both methods are frequently used in quantitative comparisons, but they share the disadvantage of a laborious experimental procedure and high costs.

The second procedure is called *label-free* quantification. This approach does not rely on any chemical modifications but directly compares the peptide signals from different samples (Wang *et al.*, 2003). To achieve this, one has to correct systematic differences in retention time between the LC-MS measurements. A number of algorithms have been established to solve this time warping or alignment problem (Prakash *et al.*, 2006; Lange *et al.*, 2007; Prince and Marcotte, 2006). For more details, we refer the interested reader to our review in Chapter 2 or the literature (Listgarten and Emili, 2005; Vandenbogaert *et al.*, 2008).

Both, labeled as well as unlabeled, quantifications rely on a detection of the peptide signals and estimation of their area as a first step. From a computational perspective, they differ only in the place where we have to search for the "partner" of a given peptide ion: either in the same LC-MS map or, after alignment, in a second map. Consequently, the algorithm that we introduce in this chapter can be applied in both scenarios.

For reasons of completeness, we mention quantification methods based on MS/MS spectra, too. One example is *spectral counting.* In this approach, we simply count the number of tandem MS spectra obtained for each peptide. This is obviously a straightforward approach. The problem is the data-dependent mode in which tandem spectra are acquired. They are usually recorded only for some of the most intense signals in each spectrum. Undersampling is always a problem, as we already mentioned in the previous chapter. Consequently, a quantification based on signal intensities in MS spectra is more sensitive since potentially all peptides in an LC-MS map can be quantified and not only the ones for which tandem spectra were obtained. Furthermore, quantifications using spectral peak intensities was shown to yield more accurate estimates of the real abundance (Old *et al.*, 2005).

Recently, another method for peptide and protein quantification has emerged. It is called iTRAQ which stands for Isobaric Tags for Relative and Absolute Quantitation and was developed by Applied Biosystems (ABI). iTRAQ is also a labeling strategy but has an important difference as compared to the methods such as ICAT or SILAC, as described above: the iTRAQ technology uses amine-specific isobaric tags and allows the simultaneous analysis of up to four different conditions in the same experiment. The tags become only visible after MS/MS fragmentation where they form reporter ions in the low mass region of the spectrum. Quantification is achieved by comparing the signal intensities of these reporter ions. iTRAQ is a promising technology for quantification and the related computational issues such as normalization and quantification of differential expression are slowly moving into the focus of the Bioinformatics community (Hill *et al.*, 2008). But it suffers from the same drawback as all quantification methods based on labeling: the mass tags are relatively expensive and the labeling procedure is laborious. In addition, many labelling technologies are restricted to peptides with matching amino acids (such as cysteine for ICAT).

Finally, due to the physicochemical limitations of the available laboratory methods, LC-MS-based proteomics usually deals with mixtures of peptides rather than intact proteins. We need to be aware that the inference of the corresponding protein abundances is a different and additional computational step (Xue *et al.*, 2006).

3.2 Preliminaries

In this section, we introduce three key concepts of our algorithm for peptide quantification: the computation of isotope distributions, the concept of averagines and the wavelet transform.

3.2.1 Computing Isotope Distributions

The naturally occurring atoms exist in different isotope states. For instance, the Carbon atom exists in two states: ^{12}C and ^{13}C with a relative abundance of 98.90% and 1.10%, respectively. The superscripts indicate the rounded mass in Dalton (Da). The lightest isotope state, ^{12}C in this case, is also called the *principal isotope state*. The *isotopes* of an element contain the same number of protons and electrons. But they differ in number of neutrons and thus in mass. Table 3.1 gives an overview of the isotope abundances of the elements most frequently occurring in peptides. Before we elaborate on the main topic of this section, we need to introduce some chemical terms and concepts that we will use during the later course of this chapter. The *average atomic mass* of a molecule is the average mass of all its naturally occurring stable isotopes weighted by their abundance. In some cases, if the mass resolution is too poor to discern single isotope peaks, the average mass will suffice. But more important in the field of mass spectrometry is the *monoisotopic mass* of a molecule. It is the sum of the masses of the lightest isotopes in the molecule.

Symbol	Isotope mass	Relative abundance (%)	avg. mass
C	12.000000	98.9	12.011
	13.003355	1.112	
H	1.007825	99.985	1.00794
	2.014	0.015	
N	14.003074	99.63	14.00674
	15.000108	0.37	
O	15.994915	99.76	15.9994
	16.999133	0.04	
	17.999169	0.20	
S	31.972970	95.03	32.066
	32.971456	0.75	
	33.967866	4.22	
	35.967080	0.02	

Table 3.1: *Isotope abundances of the atoms occurring most frequently in peptides. Adopted from de Hoffmann and Stroobant (2004).*

The *isotope distribution* of a molecule gives the masses and relative abundances of all its isotope variants. This distribution is calculated by combining the isotope abundances of the elements contained in the molecule. The isotope distribution of a molecule consisting of one element A with n_a isotope states can be computed by

$$(a_0 x^{\alpha_0} + a_1 x^{\alpha_1} + a_2 x^{\alpha_2} + \ldots + a_{n_a} x^{\alpha_{n_a}})^{m_a}$$

where a_i represents the relative abundance of the ith isotope of element A with mass α_i and m_a is the number of times the element occurs in the molecule. For a molecule consisting of the elements A, B and C, we need to compute the following polynomial

$$(a_0 x^{\alpha_0} + a_1 x^{\alpha_1} + a_2 x^{\alpha_2} + \ldots + a_{n_a} x^{\alpha_{n_a}})^{m_a} \\ (b_0 x^{\beta_0} + b_1 x^{\beta_1} + b_2 x^{\beta_2} + \ldots + b_{n_b} x^{\beta_{n_b}})^{m_b} \\ (c_0 x^{\gamma_0} + c_1 x^{\gamma_1} + c_2 x^{\gamma_2} + \ldots + c_{n_c} x^{\gamma_{n_c}})^{m_c} \quad .$$

Each power of x in the expanded polynomial represents the mass of an isotope peak with the peak intensity given by the coefficient. The number of terms to compute in this polynomial will become very large even for medium-sized molecules. To give an example, Yergey (1983) calculate that for the protein glucagon (with sum formula $C_{154}H_{224}N_{24}O_{50}S$ and molecular weight of 3485 Da) the resulting polynomial expansion contains 7.9×10^9 unique terms.

To our knowledge, the work by Carrick and Glockling (1967) is the first to deal with the efficient computation of isotope distributions. They present a program which can compute the isotope distribution for a small molecule consisting of at most three different elements (either Carbon, Hydrogen, Oxygen or Nitrogen) where each element occurs at most ten times.

Over the years, new approaches have been developed. The first algorithm that could deal with molecules of a reasonable size was developed by Yergey (1983). He uses different tricks to reduce the complexity of the problem. First, the algorithm only computes the unique isotope permutations for each element. This is in contrast to previous approaches that expand the polynomial first and then summarize like permutations (Boone *et al.*, 1970; Brownawell and San Filippo, 1982; Yergey *et al.*, 1983). Furthermore, Yergey (1983) prunes the polynomial by computing the most abundant isotopes first and discarding terms with coefficients below a user-defined threshold. Finally, Yergey (1983) exploits the fact that the absolute abundance A of each isotope permutation can be computed by

$$A = \frac{n!}{(a!)(b!)(c!)\ldots}(r_1)^a(r_2)^b(r_3)^c\ldots \tag{3.1}$$

where n is the number of atoms of the element, $r_1, r_2, r_3 \ldots$ are the abundances of each isotope and a, b, c are the numbers of the atoms in this permutation. To give an example, the expansion of the oxygen polynomial $({}^{16}O{}^{17}O{}^{18}O)^{50}$ in glucagon yields, among others, the term ${}^{16}O_{47}^{17}O_2^{18}O_1$. Its absolute abundance can be derived from

$$A = \frac{50!}{(47!)(2!)(1!)}(r_{16})^{47}(r_{17})^2(r_{18})^1 \tag{3.2}$$

by substituting into Equation 3.1 the number of oxygen atoms (50) and the numbers of the different isotopes. We can exploit this relation to reduce the number of required computations. By dividing the equations for any two permutations A_1 and A_2 by each other, we can express the abundance of one permutation in terms of the previous:

$$A_2 = A_1\frac{(a_1!)(b_1!)(c_1!)\ldots}{(a_2!)(b_2!)(c_2!)\ldots}(r_1)^{(a_2-a_1)}(r_2)^{(b_2-b_1)}(r_3)^{(c_2-c_1)}\ldots \tag{3.3}$$

where the subscripts denote the different permutations. Accordingly, Yergey (1983) compute the abundance of every first permutation and then base each subsequent permutation on the previous one. This leads to significant reduction of the number of operations required. More recent approaches (Rockwood *et al.*, 1995; Rockwood and Van Orden, 1996) apply the Fourier Transformation to convolute the isotope distributions of individual elements more efficiently. Another method (Snider, 2007) relies on Dynamic Programming.

The advantages of the above mentioned algorithms is that they can, in principle, compute isotope distributions of arbitrary mass resolution. For instance, the algorithm by Yergey (1983) outputs a list of all isotope permutations with their relative abundances above a given threshold. But for our purposes, this is not necessary: for peptide signal detection, we need a fast and flexible method that can approximate signals of different resolutions quickly, but we are not interested in isotopes with small mass differences that, in most cases, cannot be resolved by the mass spectrometer anyway. Therefore, we use an algorithm by Kubinyi (1991). It computes elemental isotope distributions by squaring intermediate distributions and prunes isotope peaks with an intensity lower than a user-defined precision. This algorithm is fast and allows us to rapidly compute isotope distributions of large molecules. Its disadvantage is that it considers integer (rounded) masses only. As an example, consider a molecule consisting of two Bromine (Br) atoms. Bromine consists of two isotopes ^{79}Br and ^{81}Br with a relative abundance of roughly 0.5 each. A simple algorithm is to split the pattern of the first atom by the distribution of the second Br atom. This leads to intensity ratios of $1 : 1 : 1 : 1$ for the peak masses $79 + 79$,$79 + 81$,$81 + 79$ and $81 + 81$. Summing equal masses leads to intensity ratios of $1 : 2 : 1$ for the masses 158,160 and 162. In principle, we can repeat this until we considered all atoms with several isotopes. But this is still time-consuming and we can reduce the computational effort even further. Note that the isotope distribution of the molecule Br_4 can be computed in a similar fashion by splitting into the distributions of Br_2 and Br_2. Accordingly, Br_8 can be computed from Br_4 and Br_4 by superimposing the corresponding distributions. Additionally, the algorithm prunes isotope peaks of low intensity in each step to reduce the overall computation time. This approach is implemented in the software library OpenMS (Sturm *et al.*, 2008) and will be used for the remainder of this thesis.

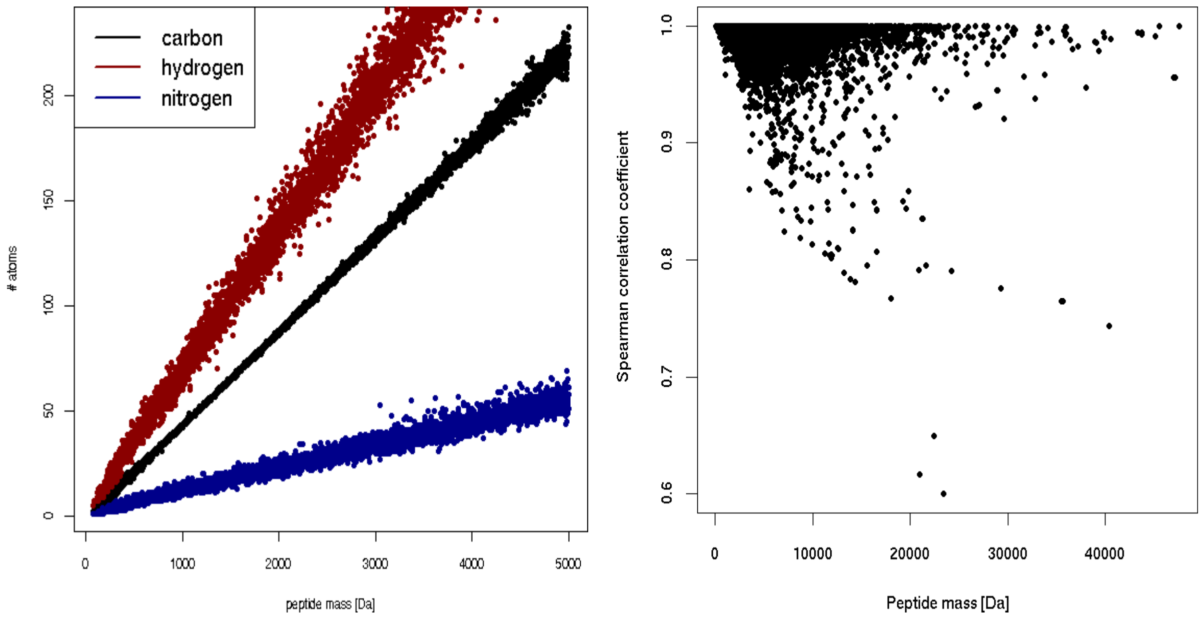
3.2.2 Averagines

In the previous section, we explained how to compute the isotope distribution of a molecule if we know its sum formula. This is important for us if we want to discriminate between mass spectral signals caused by peptides and noise. Unfortunately, if we see a signal in a mass spectrum, we usually do not know its sum formula but only the mass-to-charge ratio. Luckily, there is a way to remedy this fact.

Since we focus on peptides, we can make use of the fact that peptides are a rather homogeneous class of molecules. Figure 3.3(a) shows a plot of the number of Carbon, Hydrogen and Nitrogen atoms versus the mass of peptide sequences obtained from a tryptic digest of the SwissProt database (human protein sequences, February 2007). These atoms occur frequently in peptides and thus dominate the isotope distribution. As we can see, there is a strong linear relationship between mass and atomic composition of the peptide. Consequently, for a group of peaks in a mass spectrum, we can approximate the sum formula from the average composition of a peptide in this mass range and thus estimate the isotope pattern.

Senko *et al.* (1995a) coined the term *averagine* or *averagine model* which is simply the average composition of a peptide as outlined above. This average amino acid has the composition $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$. There are two drawbacks of this approach. First, the mass differences between isotope variants are slightly different for each element. We ignore this to simply the computations and approximate the mass differences by multiples of $\delta_{\text{av}} \sim 1.0022$ Da (Horn *et al.*, 2000). Second, we did not consider Sulfur so far. The reason for that is that Sulfur occurs very infrequently in peptides (it only occurs in the amino acids Cysteine and Methionine). But if it occurs, it has a strong influence on the isotope pattern due to the unusual distribution of its heavy isotopes (see Table 3.1).

In our averagine model, each peptide will contain a tiny amount of Sulfur which is of course not true in reality. To verify if this has a significant influence on the quality of our estimation, we compared the isotope pattern estimated from our model to the exact pattern computed from the true sum formula. We downloaded all human protein sequences from the SwissProt database (accessed February 2008) and performed an in-silico tryptic digest. For each tryptic peptide, we computed the true isotope pattern and also estimated the isotope pattern from the averagine



(a) Peptide atomic composition versus mass (b) Averagine isotope intensities versus real intensities

Figure 3.3: *(a) This plot shows the atomic composition of the tryptic peptides downloaded from the SwissProt database, sorted by their mass. There is linear relationship between peptide mass and atomic composition. This holds for the atoms most frequently occurring in peptides, only Sulfur is an exception. (b) Correlation coefficients between isotope pattern predicted by averagine and pattern computed from the true sum formula (median* 0.98*). The plot gives the difference in mass between the averagine formula and the true peptide formula (x-axis) versus the Spearman correlation (y-axis).*

of the same mass. Figure 3.3(b) shows a plot of the Spearman correlation between true isotope pattern and averagine pattern. We clearly see that this correlation is high (median 0.98), but there are a few peptides for which the correlation is relatively low. These are peptides which contain many Sulfur atoms. But even for these peptides, the lowest correlation is still ≥ 0.61. Therefore we believe that we can safely ignore the influence of Sulfur at least until new data becomes available.

To summarize, the averagine model yields a good approximation of the real signal of a peptide observed in a mass spectrum. We use this approximation twice: the mother wavelet models the isotope pattern of a peptide by an approximation of the averagine distribution. Second, we use a two-dimensional peptide template in the last step to filter for signals that were incorrectly identified as peptides.

3.2.3 Charge Determination of Isotope Peak Cluster

To summarize what has been said so far, if we encounter a group of peaks in a mass spectrum, we can estimate the peak intensities of a peptide of the same mass from the averagine composition for this mass. But so far, we ignored that the mass spectrometer in fact measures that mass-to-charge ratio and not the mass. Thus, if we want to decide if a group of peaks is caused from a peptide, we also need to estimate the charge of the compound that caused this signal.

If the mass resolution is high, the charge of an isotope pattern can easily be determined from the m/z distances between its individual peaks as we explained in Sect. 3.1.1. Hoopmann *et al.* (2007) build a histogram of the inverted m/z distances between groups of peaks in close vicinity in a mass spectrum and use the maximum value in the histogram as a charge estimate. But there also exist more sophisticated approaches.

Mann *et al.* (1989) were the first to approach the problem of charge determination. They exploit the fact that peptides in Electrospray (ESI) spectra usually appear as several charge variants (ions) and use peaks of peptide ions from the same parent mass m, but different charges z to derive the m/z of each individual ion. Their method, being among the first ESI-MS data analysis algorithms, is designed for low mass resolutions i.e. spectra in which different isotope peaks are not visible but each ion is only represented by a single, relatively broad peak. Mann *et al.* (1989) present two algorithms, called *averaging* and *deconvolution* algorithm. The averaging approach uses the m/z distance of neighboring peaks to derive an estimate for the individual charges and averages the parent mass estimate from each peak. Their deconvolution approach defines a transformation function which takes the relative intensities of all peaks in the spectrum into consideration and reaches a maximum for the parent mass which is supported by the majority of all peaks. The work by Mann *et al.* (1989) deserves credit since this lab pioneered the ESI technology and since they were among the first to tackle ESI-MS data using algorithmic means. But obviously, these methods work only if the peptide occurs as several ions, something which is not guaranteed. Their method would also fail if the spectrum is crowded and contains signals of many other compounds and peptides. There are some publications in which the authors tried to improve on these deficiencies (Henry and McLafferty, 1990; Reinhold and Reinhold, 1992), but since then the *deconvolution* approach is not longer in the focus of the mass spectrometry community.

In a more recent work, Zhang and Marshall (1998) propose an algorithm called Zscore which uses a scoring scheme for charge states to assign several possible charge states to an isotope pattern. Charge states are scored according to different criteria and depending on mass resolution. Factors that the algorithm takes into account are the logarithm of the signal-to-noise ratio and the charge states of other ions estimated to have the same parent ion mass. Each pattern is assigned the charge state with the highest score.

There are several approaches (Chen and Yap, 2008; Senko *et al.*, 1995b; Tabb *et al.*, 2006) that rely on a Fourier transform of the mass spectrum (or of a subset of the spectrum). Section 3.2.4 gives an introduction to the Fourier transform. In short, it is a method to analyze a signal for periodic components. Its application for charge determination makes sense since peaks from the same ion will appear at regular distances and thus as a periodic signal in the spectrum. Senko *et al.* (1995b) were the first to apply the Fourier transformation for charge estimation. They combine it with a Patterson charge detection routine. This routine is essentially an auto-correlation function computed for different time-lags. Tabb *et al.* (2006) compute the Fourier transformation of a small m/z window around precursor ions and compare the transformed signal to a Fourier transformation of artificial isotope pattern computed for different charge states. Windows that have a good correlation with an artificial pattern in terms of the normalized dot product are assigned the charge state of the artificial pattern. The approach developed by Chen and Yap (2008) is similar but includes a preprocessing step for candidate isotope pattern. They select a targeted peak with high-intensity from each pattern and fold (i.e. multiply) peaks to the left of this targeted peak with peaks to the right. As a consequence, symmetric peaks are kept but noise peaks are cancelled out. This improves the charge estimation for overlapping or noisy signals.

Finally, it is also possible to determine the charge state of an isotope pattern by matching it to an averagine pattern. This is coined the *matched filter* approach (Kaur and OConnor, 2006). To achieve this, Kaur and OConnor (2006) estimate the average molecular weight for a signal in a MS spectrum and compute the averagine isotope pattern for this weight. The averagine template is generated for different charge states and matched to the real signal. The charge state that results in the best fit of template to signal is used as charge estimate. The goodness of the fit is usually measured using correlation coefficients or similar measures (Kaur and OConnor, 2006).

Kaur and O'Connor (2006) and Hoopmann *et al.* (2007) compared several charge estimation algorithms on high-resolution spectra and came to the conclusion that the all algorithms described above perform similarly on high-resolution spectra, which is not surprising. The true challenge is to estimate the charge of a pattern correctly in the presence of noise, low mass resolution or if the spectrum is crowded and many signals overlap. In this case, methods based on the Fourier transformation or matched filter seem to perform best.

3.2.4 Signal Processing and Transformations

Our algorithm for peptide quantification relies on a tailored wavelet function as a first filtering step. A good way to introduce wavelet analysis is to first examine the limitations of the Fourier transformation, which is closely related to it. In this section, we briefly review the key concepts of both methods and introduce terms that will be used during the later course of this chapter. If not stated otherwise, this section is based on standard textbooks (Aldroubi and Unser, 1996; Alsberg *et al.*, 1997; Graps, 1995; Koornwinder, 1998).

We start by giving two basic definitions, that we will use in the later course of this section. The *convolution* of two measurable functions f and g is written as $f * g$ and given by

$$(f * g)(x) = \int_{-\infty}^{+\infty} f(x-u)g(u)du. \tag{3.4}$$

Second, let $L^p(\mathbb{R})$, $1 \leq p \leq \infty$ by the space of functions that are Lebesgue-integrable to the power of p. Each of these spaces is a Banach space with an associate norm $||f||_p$ which is defined as

$$||f||_p = \left(\int_{-\infty}^{+\infty} |f(x)|^p dx \right) \tag{3.5}$$

for $p < \infty$. The *Fourier Transformation* (FT) is named after the French Mathematician Jean Baptiste Joseph Fourier (1768-1830). Intuitively, we can interpret it as a decomposition of a function f into a set of basis functions. These basis functions are cosine and sine functions. The new representation of f is called the *frequency* or *power spectrum.* The Fourier transformation $F(s)$ of a function f from L^1 is given by:

$$F(v) = \int_{-\infty}^{+\infty} f(t)e^{-i2\pi vt}dt \tag{3.6}$$

for all real numbers v with $i = \sqrt{-1}$. This is one possible formulation, there are also others in terms of the angular velocity. Note that $e^{-i2\pi vt} = \cos(2\pi vt) + i\sin(2\pi vt)$, i.e. the sine and cosine basis functions are expressed in terms of the exponential function. Note that $f \in L^1$ implies that f is integrable, i.e.

$$\int_{-\infty}^{+\infty} s(t) < \infty \tag{3.7}$$

Equally important is the *inverse Fourier transformation* which results in the reconstruction of the function f and is defined as

$$s(t) = \int_{-\infty}^{+\infty} F(t)e^{-i2\pi vt}dt \tag{3.8}$$

for $v \in \mathbb{R}$. An important property of the Fourier transformation is:

Theorem 3.2.1. *(Convolution theorem) Let f and g be two functions with convolution $f * g$. Let F denote the Fourier transformation operator, such that $F(f)$ and $F(g)$ is the Fourier transformation of f and g, respectively. Then it holds that*

$$F(f * g) = F(f) \times F(g)$$

In other terms, the Fourier transformation of the convolution of f and g equals the pointwise-multiplication of the Fourier-transformed signals. This can be exploited to quickly compute convolutions. A straightforward implementation of convolution has running time $O(n^2)$. By applying the Fourier Transform (and some additional tricks), this can be reduced to $O(n \log n)$. This is particularly useful for the fast computation of isotope distributions, as we already mentioned in Sect. 3.2.1.

A disadvantage of the Fourier Transform is that it only provides information about the frequencies contained in a signal but not their localization. One solution to this problem is the *Windowed Fourier Transformation (WFT)* or *Short Time Fourier transformation (STFT)*. In short, the STFT is a Fourier Transformation computed on a short, localized window of the signal. This window is shifted along the signal such that we compute the transformation stepwise for the whole signal and by doing so, we obtain localized information on the frequency range on these local section. This window is defined in terms of a window function such that the STFT becomes effectively a convolution of the signal with this function. The STFT has the drawbacks that we need to decide on the size of the window on which it is applied and the type of the window function, each with its own advantages and disadvantages. The Wavelet theory was in part developed to cope with this problem.

The foundations of the *Wavelet transformation* were laid in 1909 by Afréd Haar. Jean Morlet, Alex Grossmann and later Stephane Mallat brought this field to the attention of the wider scientific audience in the 1980s. Jean Morlet derived the word "wavelet" from the French expression "ondelette" which means "small wave". Later, the French term "onde" (which means wave) was replaced by its English translation, resulting in "wavelet". In contrast to the Fourier transformation, the advantage of the Wavelet transformation is that it retains information about the localization of the different frequency components in the signal. Consequently, from a position in the transformed signal, we can directly map back to the the corresponding position in the original, untransformed signal.

The Wavelet transformation splits the signal into different frequency components. To do so, the signal is represented in terms of scaled and translated copies of a *mother wavelet*. These copies are called *daughter wavelets*. The Wavelet transformation exists in a continuous and a discrete version. The *Continuous Wavelet transformation (CWT)* W_ψ of a function $f \in L^2(\mathbb{R})$ is given by

$$W\psi[f](a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x)\psi(\frac{t-b}{a})dx \tag{3.9}$$

where $a \in \mathbb{R}\backslash\{0\}$ and $b \in \mathbb{R}$ are the scale and position parameter, respectively. Intuitively, we shift the mother wavelet along the signal and compute the transformation for each position. We repeat this for different scales to obtain representations of the signal at different resolutions. Equation 3.9 can also be cast as a convolution of the signal with the translated and dilated wavelet. For practical applications, it is desirable to have a mother wavelet $\psi \in L^2(\mathbb{R})$ with the following properties:

- **Compact support** The support of a function is simply the set of values on which it is not zero. The support is compact if this set of values is closed and bounded. Thus, the requirement of compact support means that the mother wavelet decays rapidly towards infinity and is $\neq 0$ only on a short interval.

- **Zero mean** The average value $\int \psi(t)\ dt$ is zero, i.e. $\int_{-\infty}^{+\infty} \psi(t) dt = 0$. The wavelet must thus be oscillatory.

The wavelet transformation also exists in a discrete version. Discrete means in this context that we sample the wavelet only for discrete scales and positions. Thus, in the *Discrete Wavelet Transform (DWT)*, the wavelet ψ is modified to

$$\psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}} \psi \left(\frac{t - k\tau_0 s_0^j}{s_0^j} \right) \tag{3.10}$$

where j and k are integers, $s_0 > 1$ is called the dilation step and τ_0 is called the translation factor. A common choice for s_0 is 2 such that the signal is sampled at every step of 2.

3.3 Previous Work

There are several algorithms for the detection of peptide features in mass spectra. The first step is usually a *seeding* step: the algorithm tries to identify prominent peaks in the map that could be part of a peptide signal. Feature detection algorithms mostly differ in the criteria they use to select these seeds. After this, most algorithms separate the seeding points from the surrounding background and compute a preliminary bounding box.

After this, all algorithms perform a *refinement* step: all data points in the previously computed bounding box are compared to an averagine model, as described in Sect. 3.2.2. Again, the algorithms differ in the complexity of their averagine model. Most algorithms do not take the elution behavior of the isotope peaks into consideration and some even apply only a crude approximation to the true isotope pattern. Quantification algorithms also differ in the mass resolution and accuracy they expect: some require highly resolved spectra, others perform well only on low resolution data. In the remainder of this section, we will review the state of the art of LC-MS feature detection algorithms. We will however take only algorithms into consideration that are freely available for academic purposes i.e. no commercial software or algorithms for which no free implementation exists.

3.3.1 Seeding

There are several possible criteria for seed selection. A straightforward approach is to sort all data points in the LC-MS map and to check all points above a given intensity threshold (Gröpl *et al.*, 2005). This approach has also been coined as *intensity descent* (Chen *et al.*, 2006). This strategy is straightforward to implement but becomes prohibitively slow if the LC-MS map is complex and contains many noise signals. In this case, an intensity-based seeding method will select many noise signals or non-peptidic signals as feature candidates and perform many unnecessary refinement steps or even require manual inspection to yield good results. Thus, it makes sense to take more criteria than the intensity of a signal into consideration. Katajamaa *et al.* (2006) compute various higher-order statistics for each seed, such as signal-to-noise ratio, consistency across several spectra and intensity difference to neighboring points. Only seeds that pass user-defined thresholds in all these criteria are considered valid seeds. The approach by Chen *et al.* (2006) is similar in spirit, but in contrast to Katajamaa *et al.* (2006) they interleave refinement and seeding step. In short, they check groups of high-intensity points for correlation with an averagine model first, remove them from the spectrum and only then proceed with lower intensity signals. This leads to fewer false positives. Du *et al.* (2007) also perform an intensity descent but also take more sophisticated peak characteristics into account. First of all, their algorithm does not operate on data points but groups points into peaks and only considers peaks of a user-defined minimum width. The algorithm also considers only peaks that occur at consistent m/z positions across at least 5 spectra. It iteratively groups adjacent peaks of high-intensity into a cluster, fits an averagine model for refinement and then proceeds until all peaks are processed.

An LC-MS map can be represented as a two-dimensional image if we interpret m/z and retention time as the two dimensions and represent the intensity by a color code. It is therefore an evident idea to use image processing methods for LC-MS feature detection. Leptos *et al.* (2006) apply the Watershed transform (Vincent and Soille, 1991) which is another intensity-based seeding approach. The Watershed transformation performs a segmentation of the LC-MS map. The resulting segments are disjoint and homogeneous regions in the map. The Watershed transformation comes from the field of digital image processing and is usually defined in terms of discrete and equally spaced points. Consequently, Leptos *et al.* (2006) resample the LC-MS map to achieve a rectangular grid. In general, there are two classes of algorithms to compute the Watershed transform: algorithms based on distance functions or on recursion. Leptos *et al.* (2006) do not explain which approach they use. In any case, the watershed transformation is fast but the resampling might decrease the resolution of the data. Furthermore, the watershed transformation does not take the shape of the signal into consideration which is an important and reliable information in the case of isotope pattern.

The THRASH (Thorough High-Resolution Analysis of Spectra by Horn) algorithm (Horn *et al.*, 2000) uses a seeding approach based on a histogram of the signal intensities. Apart from that, this algorithm is similar to the previous ones. But it is noteworthy since it was the first published LC-MS feature detection algorithm. Unfortunately, its first implementation is not longer available. The software DECON2LS implements the THRASH algorithm and is available from the NCRR Proteomics Resource at the Pacific Northwest National Laboratory, Richland, US.[1]. The algorithm was also re-implemented and modified by Hoopmann *et al.* (2007). In short, this method assumes that the background signal over a given m/z range will be acquired more frequently than the signal of the isotope peaks. Both algorithms build a histogram of the signal intensities for a small m/z range, around 25 Th. The background intensity is estimated as the intensity with the highest frequency. The noise is estimated by Full-Width-at-Half-Maximum (FWHM) of the smoothed histogram, and the signal-to-noise ratio (S/N) is calculated for each signal point (I_p) using

$$S/N = \frac{I_p - I_b}{\text{FWHM}}$$

where I_b is the background intensity. Peaks surpassing a user-defined S/N threshold (usually around $2-3$) are chosen as seeds.

The software tools MSINSPECT (Bellew *et al.*, 2006) and SPECARRAY (Li *et al.*, 2005) both use the wavelet transformation to filter the raw spectra and later search for local-maxima to identify peaks. The algorithms differ in the type of mother wavelet they use. Bellew *et al.* (2006) apply the Marr-Wavelet (also called Mexican Hat) whereas Li *et al.* (2005) use the Symmlet8-wavelet. The wavelets help to suppress signals without the typical shape of a MS peak and thus

[1] http://ncrr.pnl.gov/software/

reduce the number of seeds to be considered. Unfortunately, the respective publications do not give further details on the feature discovery process.

Finally, the software Superhirn (Mueller *et al.*, 2007) implements a fast feature detection routine which is an improved version of the pattern matching method previously published in Gay *et al.* (2002). In short, the algorithm starts by fitting an averagine template to peaks with high intensity in each spectrum. Peptide features are identified by searching for local minima in the error function, i.e., in the squared absolute distance between averagine template and spectrum. They use several heuristics to improve the running time of their algorithm. As an example, they compute the error function for larger bins of spectra and region with low noise content first. After well-resolved features are found, they remove the corresponding peaks from the spectra and re-compute the error function to find lower intensity and overlapping features.

3.3.2 Refinement

All these approaches result in a set of potentially interesting signals in the LC-MS map which might be caused by a peptide. In a next step, these seeds are filtered for false positives. To do so, we fit a theoretical model of a peptide signal to the peaks in the neighborhood of each seed and determine the goodness of this fit, usually using a measure of correlation of peak intensities predicted by the model and the data. Signals with a poor fit to the theoretical predictions are discarded. Almost all algorithms introduced above use the averagine model as introduced by Senko *et al.* (1995a). We described this model in Sect. 3.2.2. Current approaches mainly differ in the implementation and the method of charge determination. Table 3.3.2 gives an overview. In the following, we will highlight these differences.

Name	Reference	Seeding strategy	Charge determination
AID-MS	Chen *et al.* (2006)	intensity descent	interpeak spacing
Hardkloer	Hoopmann *et al.* (2007)	histogram	m/z distance
lcms2d	Du *et al.* (2007)	intensity descent	variable selection
MapQuant	Leptos *et al.* (2006)	watershed	averagine model
msInspect	Bellew *et al.* (2006)	Marr Wavelet	averagine model
MZmine	Katajamaa and Orešič (2005)	intensity descent	averagine model
SpecArray	Li *et al.* (2005)	Symmlet8 Wavelet	m/z distance
Superhirn	Mueller *et al.* (2007)	error function	averagine model
THRASH	Horn *et al.* (2000)	histogram	FT + Patterson

Figure 3.4: *Overview of quantification algorithms including references.*

Noteworthy is the software packages MapQuant since it includes a parameter of the peak shape into its averagine model. This tool models the peak shape using a Gaussian distribution. By changing the peak widths, these algorithms can account for different mass resolutions. This is also a feature of our own algorithm. Unfortunately, MapQuant reads only proprietary file formats and is apparently not longer maintained. All other algorithms do not take the peak shape into consideration but reduce the raw data points to peaks and compare peak intensities to averagine intensities.

An interesting feature of Hardkloer is its ability to model non-standard isotope pattern. The user can define averagine amino acids with non-standard composition to account for modified or isotope-enriched peptides. Obviously, this makes only sense for high-resolution data.

Finally, the software Hardkloer uses a simplified averagine model which does not compute the polynomial as we described it in Sect. 3.2.1. Their model is based on a Poisson distribution which is fast to compute but results in a less accurate model (Bellew *et al.*, 2006).

3.4 Our Contribution

A typical proteomic sample consists of several thousand peptides and clinical studies often consist of hundreds of samples. It is therefore desirable to detect and quantify all peptides in a sample as quickly as possible. In particular, the matching of the extracted regions to an averagine model takes time and might even require a manual validation of the signal. An efficient algorithm that is suitable for real-world applications should thus aim for few candidate signals and therefore few fitting attempts. Nevertheless, if seeds or regions in the map are chosen based on their intensity alone, without taking the expected shape into consideration, we will obtain a lot of false positives, many of which will simply be caused by chemical noise or contaminants of the sample. This in turn results in many unnecessary fitting attempts, a slow quantification, and more false positive peptide signals which will hamper a further down-stream analysis.

We present our algorithm, called SweepWavelet, which introduces a model-based approach to peptide quantification. This algorithm was developed in a joint project with researchers at the University of Saarbrücken, Germany. The key idea of our approach that we know what we are looking for in an LC-MS sample (i.e. peptides), and since we know how to model their distinctive patterns in a mass spectrum, there is no need for a global segmentation based solely on peak intensity. It makes more sense to rely on a model-based approach and to use our knowledge about the isotope pattern of an average peptide.

In Schulz-Trieglaff *et al.* (2007), we have shown that we can accurately and quickly quantify even low abundance peptides using a mother wavelet which mimics the distribution of isotope peak intensities. Peptide signals occurring in adjacent spectra are joined using a sweep-line algorithm inspired by computational geometry. Furthermore, we use an averagine which contains parameters for peak width and elution behavior of the isotope pattern such that we can separate peptide signals accurately from the noise. In this chapter, we provide details on our model of isotope peak intensities and how we use the redundant information in adjacent spectra to estimate monoisotopic mass and peptide charge state. The next chapter provides a detailed comparison of our approach with other popular programs on selected benchmark data sets.

Note that we do not commit ourselves to a specific labeling technique or instrument type apart from assuming a standard LC-MS setup. We focus on the precise detection and quantification of the mass spectral intensities of peptide ions. But our approach is more flexible and can be applied to various scenarios, such as accurate mass and time (AMT) tag approaches (Silva *et al.*, 2005; Smith *et al.*, 2002) or for peptide mass fingerprinting (Hussong *et al.*, 2007).

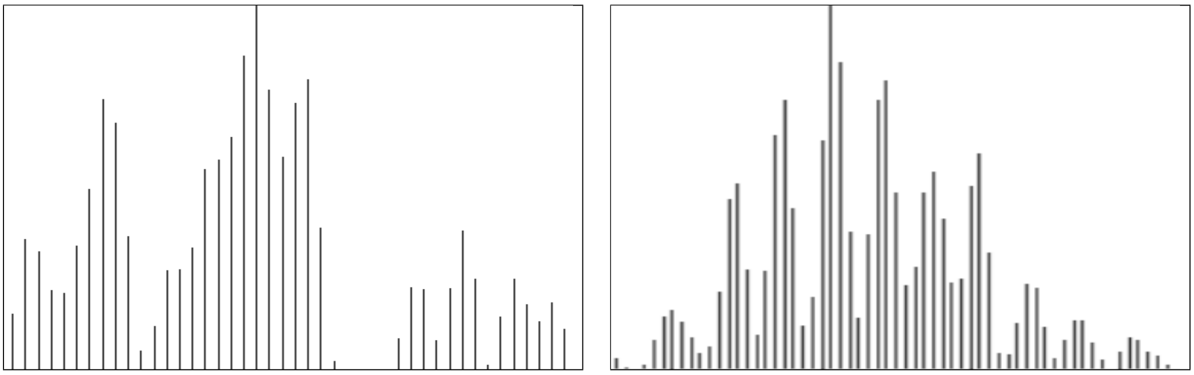
3.5 Algorithm

Our approach is divided into the following stages: preprocessing, seeding (detection of candidate signals), clustering of adjacent signals and filtering. During the seeding stage, we search for groups of candidate peaks resembling an isotope distribution. Then, we join isotope peak patterns in neighboring spectra using an approach inspired by sweep-line algorithms from computational geometry and finally, we filter the obtained peak groups by matching them against an averagine model.

3.5.1 Preprocessing

Our algorithm is implemented as a part of a larger collection of computational tools for the analysis of mass spectra (Kohlbacher *et al.*, 2007). These tools, such as algorithms for baseline subtraction or noise removal, can easily be combined with our approach. However, our algorithm does not depend on any previous filtering or noise reduction of the spectra.

We perform a simple preprocessing of the LC-MS map to improve the quality of the seeding phase: for each data point in each spectrum, we check whether it occurs again in the following spectra. If this is the case, we add a bonus to its intensity to reflect our higher confidence that this point represents a meaningful signal. This bonus is the sum of the intensities of the signals at the same mass (within a small tolerance) in the next five spectra. We use a small m/z tolerance dependent on the instrument but usually around 0.01 m/z for high-resolution spectra.



(a) Single spectrum

(b) Aligned and projected spectra

Figure 3.5: *(a) A single spectrum from the middle of the elution curve of a feature. (b) The same feature after alignment and projection of its* 7 *spectra.*

Note that we do not use this increased intensity during quantification but only during the wavelet-based filtering of the spectra. This preprocessing step helps to raise isotope patterns of low intensity above the noise level and implicitly assigns a penalty to noise signals which do not occur in several spectra at similar positions.

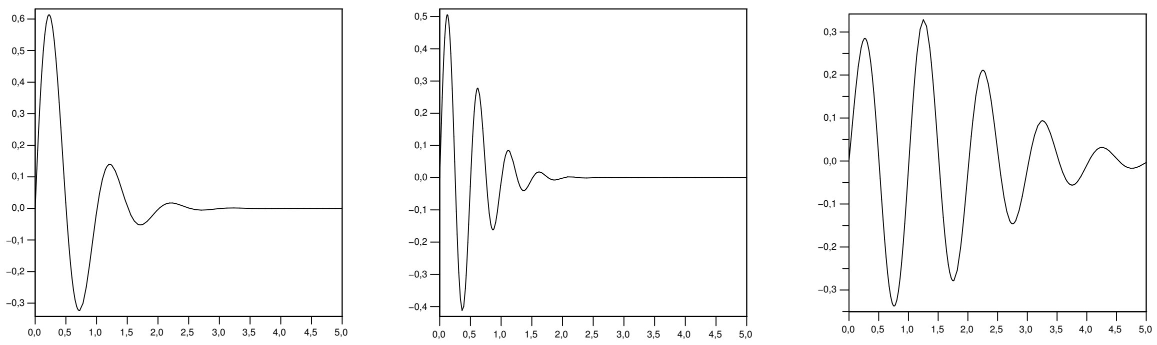
3.5.2 Peptide Candidate Detection using Wavelets

As described above, the aim of this step is to detect candidate regions (seeds) in the spectra of a LC-MS map, regions that are likely to contain isotopic peaks caused by a peptide. The key ingredient of this step is a wavelet function that models the typical shape of an isotope pattern. Note that we do not perform a full-scale wavelet analysis of the mass spectrum. We can imagine the wavelet transformation as a filtering step. Our wavelet function suppresses signals that do not match its shape but raises matching signals high above the noise level.

Since an isotope pattern depends on the mass and charge of the ion causing it, our wavelet also takes this fact into consideration and is a function of mass and charge.

3.5.3 The Isotope Wavelet

We use a tailored mother wavelet to filter mass spectra for isotope pattern Since then, the implementation of the isotope wavelet was further improved Hussong *et al.* (2007). They applied it to MALDI spectra and for peptide mass fingerprinting. Here, we present its application to LC-MS data and quantification. As outlined above, an isotope pattern consists of a sequence of



$m/z = 500, z = 1$ $m/z = 500, z = 2$ $m/z = 2000, z = 1$

Figure 3.6: *The isotope mother wavelet for different masses and charges. These plots are by courtesy of Rene Hussong, ZBI Saarbrücken, and were already used in Schulz-Trieglaff* et al. *(2007).*

peaks, whose intensity ratios depend on the mass of the peptide ion causing it. We use a sine function to model the sequence of isotope peaks and truncate the sine after a certain number of oscillations to achieve compact support. To model the decaying peak intensities, we multiply it with the continuous equivalent of a Poisson distribution. The resulting building block of the isotope wavelet is:

$$\psi_i(t,\lambda,\mu) := \frac{\sin(2\pi\mu t)\cdot\exp(-\lambda)\cdot\lambda^{\mu t}}{\Gamma(\mu t+1)} \tag{3.11}$$

$$\psi(t,\lambda,\mu) := \theta(t)\cdot\psi_i(t,\lambda,\mu) = \begin{cases} \dfrac{\sin(2\pi\mu t)\cdot\exp(-\lambda)\cdot\lambda^{\mu t}}{\Gamma(\mu t+1)} & \text{if } t \geq 0 \\ 0 & \text{else} \end{cases} \tag{3.12}$$

where $\theta(t)$ denotes the unit step function. To achieve compact support, we truncate the function after some cut-off t'. But to obtain a wavelet, we need a function with zero mean, as stated in Sec. 3.2.4. Thus, we simply subtract the mean value of the function obtained so far. This yields:

$$\psi(t,\lambda,\mu) = \begin{cases} \frac{\sin(2\pi\mu t)\cdot\exp(-\lambda)\cdot\lambda^{\mu t}}{\Gamma(\mu t+1)} - \int_0^{t'} \psi_i(t,\lambda,\mu)\, dt & \text{if} \quad 0 \leq t < t' \\ \\ 0 & \text{else} \end{cases} \tag{3.13}$$

Besides t, the time/place variable, the wavelet has two additional parameters: λ and μ. λ is mass-dependent and reflects the ratios between different peaks of the isotope pattern. μ is the charge of the pattern we investigate. Therefore, it controls the periodicity of the function as well as the support of the wavelet. Figure 3.6 shows plots for three different isotope wavelets for different charges and masses.

Since the shape of the isotope patterns depends on the mass as well as on the charge of the considered peptide, we use several wavelet functions. While we adopt the mass parameter automatically during the computation of the transform, we need to compute a different wavelet function for each charge state. Therefore, if we assume peptides to be at most x-fold charged, we have to compute x transformed versions of each spectrum. The advantage of this approach is that we obtain an estimate of the charge of the isotope pattern for free, since the mother wavelet with a charge matching the charge of the ion will yield the best correlation. With the isotope wavelet, we can detect regions of interest in the spectrum by searching for local maxima in the

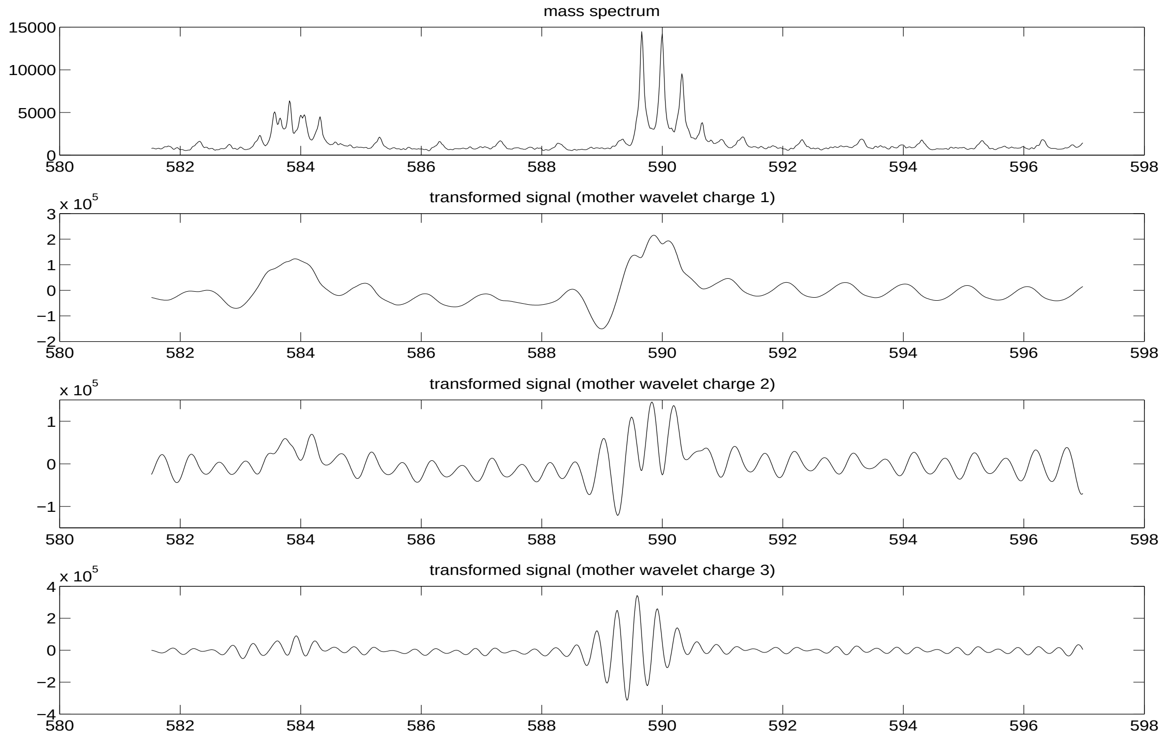


Figure 3.7: *A mass spectrum and its wavelet transformation for mother wavelets of charge one to three. The strong oscillating part of the transformed signal is a peptide of charge three, the remaining peaks are non-peptidic compounds or noise.*

transform. A 'real' isotope pattern will lead to a chirp-like signal in the wavelet transform, since each of its peaks will lead to a resonance with the wavelet, and we make use of this special shape to improve the specificity of our approach. The positive parts of the wavelet closely resemble an isotope pattern, while its negative parts 'punish' high spectral intensities in between peaks. Hence, a high correlation of a signal with the isotope wavelet occurs if the signal resembles an isotope pattern, while deviations from the assumed pattern will lead to a loss of correlation. In the case depicted in Fig. 3.7, the peak cluster at 590 Th represents a peptide of charge 3 whereas the signal at 584 is a non-peptidic compound. The region between 598 Th and 600 Th contains several noise peaks.

Due to the design of the wavelet, we only need to compute one scale of the wavelet transform, i.e., compute the correlation integral of the isotope wavelet with the mass signal. While at first glance this seems to require $\mathcal{O}(n^2)$ operations where n is the number of points in the mass spectrum, the real runtime is actually much smaller: since the mother wavelet has finite and typically small support that is independent of the length of the mass signal, we can compute the transformation in linear time.

3.5.4 Scoring of Signals

Sequences of isotope peaks in the mass spectrum result in intervals with strong oscillations in the wavelet-transformed signal as we see in Fig 3.7. We compute a score for all disjoint subintervals in the transformed spectrum based on an F-statistic:

$$F_{(n-1,n-1)} \sim \frac{\hat{\sigma_1}}{\hat{\sigma_2}} \tag{3.14}$$

This corresponds to a two-sample F-test for the equality of variance of two samples. Here $\hat{\sigma_1}$ is the variance of the subinterval in the transformed signal and $\hat{\sigma_2}$ is the variance in an equally sized interval representing the baseline variance in the wavelet transform. n is the size of the subinterval.

There are many ways conceivable in which a baseline variance of the wavelet-transformed mass spectrum can be computed. One could use the variance from an empty spectrum e.g. a spectrum without peptide features or other signals. But such "empty runs" often require extra manual work and are not always available. Alternatively, one could try to estimate the background variance from the whole spectrum. Note that $\hat{\sigma_1}$ and $\hat{\sigma_2}$ in Eq. 3.14 have to be statistically independent. For reasons of simplicity we decided to test against the variance of an interval with the same size at the end of the spectrum since this part of the spectrum is unlikely to contain isotope peaks or other signals of interest. We tested our assumption of statistical Independence for several LC-MS maps. A χ^2 test for the Independence of signal variance and baseline variance resulted in p-values of < 0.01. Of course it is also possible to perform the F-test against user-specified values for $\hat{\sigma}_2$ and n.

In practice, a typical spectrum can contain several thousands of signals causing a high variance in the wavelet-transform. The testing of hypotheses at this level of multiplicity will necessarily lead to many false positives. Therefore we use a procedure developed by (Benjamini and Hochberg, 1995) to circumvent this problem. This approach relies on a control of the *False Discovery Rate* (FDR), defined as the number of falsely rejected null hypotheses (i.e. false positives). The FDR provides a cutoff for oscillations in the wavelet transform. We sort the unadjusted p-values $p_1, p_2 \ldots p_n$ and choose j^* such that $j^* = \max\{j : p_j \leq \frac{j}{n}\,\alpha\}$ where n is the number of tests performed. To control the FDR at level α, we discard all intervals with p-values $p_i > p_{j^*}$.

A similar approach was used in (Tan *et al.*, 2006) to test for regions of significance in low-resolution SELDI spectra. They performed an analysis of variance (ANOVA) directly on the mass spectrum and test for significant differences against spectra obtained from blank runs. We also considered the use of a blank spectrum but testing against the end of each spectrum yielded comparable results (data not shown).

3.5.5 Combining Patterns in Adjacent Scans

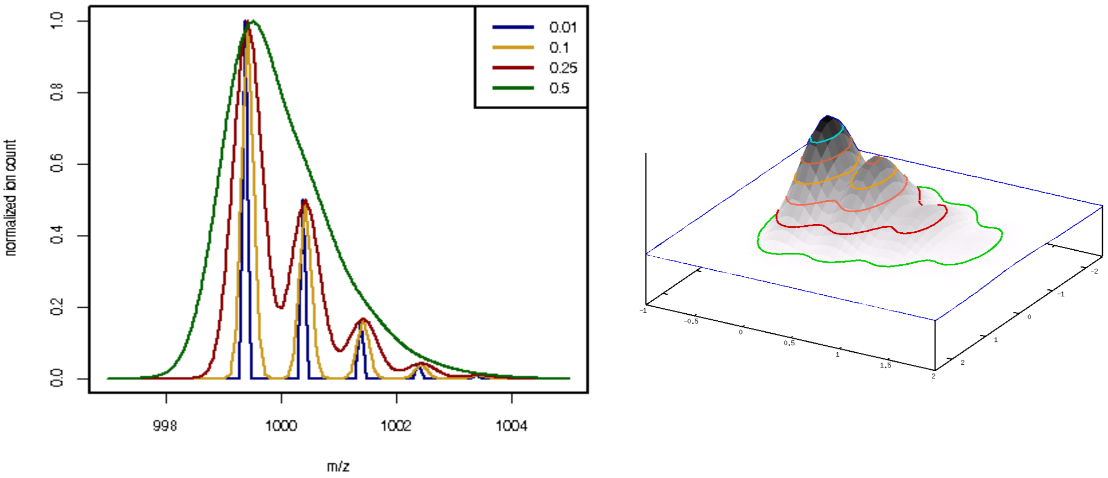
We use an approach inspired by the sweep-line algorithm from computational geometry to join isotope patterns in adjacent spectra that are likely to be caused by the same peptide ion. The sweep-line algorithm is a general paradigm which is used by many algorithms. The idea can be illustrated by an imaginary line sliding over the input data. Specific events are triggered as the line passes over interesting points of the scene, such as the beginning or the end of an object. The algorithm keeps track of the objects it encounters using a dynamic data structure.

In our case the interesting objects are adjacent and possibly overlapping isotope patterns. Hence, we sweep across the LC-MS map and use the isotope wavelet to detect the starting positions of isotope patterns in each spectrum. That is, we apply the wavelet transformation to each spectrum in order of increasing retention time. A significant signal in the wavelet transformation triggers an event. We check if the previous spectrum contains a pattern at a similar mass up to a small tolerance. The predicted monoisotopic masses of each pattern in each spectrum are kept in a tree data structure.

Since we allow for overlapping signals, we store a first guess of the monoisotopic m/z and the charge state z for each pattern. The algorithm joins patterns located in adjacent spectra and computes a bounding box around the region which is occupied by them in the LC-MS map. The bounding box is closed when no more events arise extending it. For each box we determine the monoisotopic mass and charge by voting across all spectra in which we detected the pattern.

In this way we obtain a first estimate for monoisotopic mass-to-charge and charge state, as well as a bounding box, for all peptide signals in the LC-MS map. However, this set is still likely to contain false positives. Therefore we match the extracted signals against a pre-computed template in the next step and by doing so, filter out incorrectly extracted signals.

3.5.6 Filtering using a Peptide Template



(a) Averagine model

(b) Full two-dimensional model

Figure 3.8: *Left: The averagine model of our algorithm with peak widths (standard deviations for the Gaussian distribution) ranging from* 0.01 *to* 0.5*. Right: the full two-dimensional peptide template including a Gaussian model for the elution profile. We use this model to filter out false-positive signals. The right plot is by courtesy of Clemens Gröpl.*

For each potential peptide signal identified in the two previous steps of the algorithm, we fit a two-dimensional template to its region. In contrast to the isotope wavelet, this template or model also takes the typical elution time of a peptide into account. The relative intensities and distances of its peaks along the m/z axis are given by an averagine isotope distribution for the given mass region. Moreover, we model the imprecision of the mass analyzer by a Gaussian distribution with variance σ^2. In this way we can account for different MS instruments and different mass accuracies. Hence, our model for the isotope distribution of a peptide is

$$\phi(m) = \frac{A}{\sqrt{2\pi\sigma^2}} \sum_{i=0}^{i_{\max}} a_i(m_0) e^{-(m-m_0-i\delta_{\text{av}})^2/(2\sigma^2)}\,, \tag{3.15}$$

where m_0 = monoisotopic mass, $a_i(m_0)$ = relative abundance of i-th isotopic peak of a peptide with monoisotopic mass m_0, $i_{\max}$ = last isotopic peak considered, A = area under curve, and δ_{av} is the distance between consecutive isotope peaks. Figure 3.8(a) shows the m/z part of our template for different settings of σ^2. We usually keep the width σ^2 of the isotope peaks fixed for a given LC-MS experiment. It can easily be determined by visually inspecting a couple of isotope peaks in the LC-MS. We assume the peak shape to be Gaussian and the Full-Width-At-Half-Maximum (FWHM) of a Gaussian function is given by $2\sqrt{2\ln 2}\sigma$, where σ is the standard deviation of the Gaussian.

In many cases, a simple standard Gaussian function can be used to describe the elution profile of a peptide. Nevertheless, our template also accounts for asymmetric peak shapes (due to fronting or tailing effects) and can incorporate more versatile profiles such as the exponentially modified Gaussian (EMG) (Grushka, 1972). Several studies have shown that it provides a good fit for chromatographic peaks in reversed-phase chromatography (Li, 2002; Naish and Hartwell, 1988). It is defined as

$$\mathrm{EMG}(x) = \frac{ac\sqrt{2\pi}}{2d} \exp\left(\frac{b-x}{d}\frac{c^2}{2d^2}\right)\left(\frac{d}{|d|} - \mathrm{erf}\left(\frac{b-x}{\sqrt{2}c} + \frac{c}{\sqrt{2}d}\right)\right) \tag{3.16}$$

where b is the standard deviation of the Gaussian part, d is the expected shift of the exponential modifier, a the amplitude and c the center. For $d < 0$, we obtain a fronted peak and for $d > 0$, the peak is tailed.

Depending on the type and quality of the chromatographic separation, the user can either choose an elution profile or let the algorithm iterate over a set of function and choose the one which gives the best fit. Figure 3.8(b) shows the full two-dimensional model consisting of averagine and Gaussian elution model.

3.5.7 Peptide Signal Model and Bounding Box

Using our two-dimensional model, the candidate regions from the seeding phase are evaluated as follows. In a first step, we compute our averagine model only for the mass and charge as predicted by the isotope wavelet. We project the candidate region on its m/z axis and align the averagine template with the signal points in the bounding box extracted by the sweep-line algorithm. Then, we determine the goodness of fit using the Spearman's rank correlation between the intensities predicted by our template and the intensities in the data. Spearman's rank correlation is a non-parametric measure of correlation. This means that it makes less stringent assumptions about the underlying data distributions. It is similar to the Pearson correlation but the data, in our case the template and data intensities, are converted to ranks before computing the correlation coefficient. To do so, the intensity values are sorted and each intensity is replaced by its rank R in the sorted order. The rank correlation coefficient $r^S_{x,y}$ is then defined as

$$r^S_{x,y} = \frac{\sum_{i=1}^n (R_i^M - \mu)(R_i^D - \mu)}{\sqrt{\sum_{i=1}^n (R_i^M - \mu)^2}\sqrt{\sum_{i=1}^n (R_i^D - \mu)^2}} \tag{3.17}$$

where R_i^M and R_i^D are the ranks of model and data points, respectively, and μ is the mean of the ranks which is $\frac{1}{2}(n+1)$.

Fitting the EMG function to the data is more difficult. We use an approach developed in Grunert (2008) and fit the EMG using the Levenberg-Marquardt algorithm (LMA) (Levenberg, 1944; Marquardt, 1963). The LMA is a least-squares curve fitting algorithm and is a widely used algorithm described in standard textbooks (Nocedal and Wright, 2006). In short, we project the candidate region on its retention times axis. Given the set of rt_i and intensity (it_i) values, the algorithm determines a set of parameters β for the EMG such that the sum of the squared deviations E between data intensities and function becomes minimal:

$$E(\beta) = \sum_{i=1}^{m} [\mathrm{it}_i - \mathrm{EMG}(\mathrm{it}_i, \beta)]^2. \tag{3.18}$$

The LMA is an iterative gradient-descent algorithm. In each iteration step i, the parameter vector β_i is replaced by a new estimate β_{i+1}. The standard gradient descent approach is to update the parameters according to

$$\beta_{i+1} = \beta_i - \lambda \nabla f(x) \tag{3.19}$$

where $\lambda \nabla f(x)$ is the scaled gradient. This is a straightforward method but it has several deficiencies. Among other issues, it takes the counter-intuitive approach of taking large steps if the gradient is large but only small ones if the gradient is small. This leads to a poor convergence behavior. The idea of the LMA is to incorporate second-order derivatives into the update step :

$$\beta_{i+1} = \beta_i - (H + \lambda \mathrm{diag}(H))^{-1} \nabla f(x) \tag{3.20}$$

where H is the Hessian matrix evaluated at β_i, λ is the scaling factor and $diag(H)$ is the diagonalized Hessian. The Hessian matrix contains all second-order partial derivatives and is a measure of the curvature of f. Thus, Eq. 3.20 implies a large step in the direction with low curvature (i.e., an almost flat terrain) and a small step in the direction with high curvature (i.e, a steep incline). The first formulation of the algorithm (Levenberg, 1944) using the identity matrix instead of the diagonalized Hessian in the update rule. The current version was suggested by Marquardt (1963) and has the advantage that each dimension of the gradient is scaled by the curvature. This leads to larger steps into the directions where the gradient is smaller.

After aligning the averagine template for m/z and fitting the EMG for the rt dimension of the region, we computed the goodness of the fit using Spearman's rank correlation. If the correlation is below a threshold, we re-compute the averagine model for a range of charge states and if this results in a sufficiently good fit, we keep the extracted region. The monoisotopic mass is estimated from the theoretical isotope distribution, and the coordinate in retention time is taken as the centroid of the (modified) Gaussian distribution.

The threshold for goodness of fit was estimated by inspecting LC-ESI-MS data from a BSA digest containing up to three charge variants per BSA peptide. We tried different values for the minimum correlation required to report a match between template and signal until all BSA peptides were detected by our algorithm. This was the case for a threshold of 0.6, and we kept this threshold for all data of similar resolution and quality.

There are several imaginable ways how one might measure the goodness of the match between model and peptide signal. Apart from a correlation-based measure, the literature contains various other approaches such as the Kullback-Leibler divergence (Bellew *et al.*, 2006), the dot-product on centroided spectra (Hoopmann *et al.*, 2007) or the Pearson correlation coefficient (Chen *et al.*, 2006).

Finally, we perform an adjustment of the feature bounding box: starting from the data points on the convex hull of the feature, we examine points in their neighborhood which are not part of the feature. If the real intensity and the intensity predicted by our model of any of these data points exceeds a user-defined threshold, we include the point into the feature and extend the bounding box accordingly. This helps in cases when the initial estimate of the bounding box was too small.

3.5.8 Separation of Overlapping Isotope Patterns

In complex samples such as serum or whole cell digests, co-eluting peptides might cause overlapping isotope patterns. To achieve a correct quantification, these overlapping signals need to be separated and, for each data point in this area, we need to decide to which peptide ion it belongs.

We propose a greedy iterative approach to solve this problem. If we detect overlapping isotope patterns in the wavelet transform, we join them independently during the sweep-line stage and allow for overlapping bounding boxes. Subsequently, we fit several templates to this region, starting with the template corresponding to this highest scoring mass and charge combination. We flag points having a good correlation with the template as used and continue with the next template until no more hits in this region of the LC-MS map remain. In this respect, our approach is similar to the one presented by Horn *et al.* (2000).

3.6 Computational Experiments

In this section, we present two computational experiments to support the claims made in this chapter. First, we show that we can efficiently filter mass spectra for isotope signals using the isotope wavelet. Second, we show that we can accurately determine the charge state of the detected isotope pattern using our averagine model. Finally, we present a case study of the quantification capabilities of our algorithm using clinical LC-MS data. More detailed evaluations will follow in the next chapter.

3.6.1 Filtering for Isotope Pattern

As we outlined in Sect. 3.3.1, all algorithms perform a *seeding* step in which they try to identify potential interesting signals (seeds) and then perform a *refinement* step in which they fit an averagine model to the signal. They signal is discarded if the quality of the fit is poor. This helps to reduce the number of false positives.

Most algorithms do not take the shape of the signal into consideration but use only simpler criteria such as the intensity. If we select the seeds in the LC-MS map only based on their intensity, we refer to this approach as *intensity descent*. The disadvantage of this approach is that many unnecessary refinement steps are performed since many signals selected will not be peptides but can be any signal or even noise, especially when we approach lower intensities just above the noise. But lower intensities cannot be skipped, since there might be a significant number of signals in this region. Furthermore, the refinement step is computationally expensive since it involves fitting a complex model with a difficult fitness landscape. That is why we would like perform a model fit only if it is really necessary. An intensity descent will also lead to more false positives since many seeds will pass the quality threshold by chance alone without being true peptide features.

To verify if our model-based approach using the isotope wavelet does indeed result in a reduced number of model fitting attempts, we implemented a straightforward intensity descent algorithm which uses the same two-dimensional model as our algorithm, SweepWavelet. We compared the number of seeds for both methods on data sets of similar complexity recorded on different mass spectrometers. Our hypothesis is that our algorithm SweepWavelet will always need less seeds (and thus fitting attempts) to obtain the same number of features than an intensity descent approach. The difference in seeds will increase for larger LC-MS maps recorded on newer instruments since these instruments can resolve more peptide features.

We obtained a set of four LC-MS maps recorded on different mass spectrometers. Table 3.2 gives an overview. The complexity of the samples is roughly comparable but size and resolution of the maps increase with the sample number. The first and second LC-MS map were recorded using published protocols (Mayr *et al.*, 2006; Whitehead *et al.*, 2006). The third and fourth maps are by courtesy of Dr. Mark Robinson at the Walter and Eliza Hall Institute in Melbourne,

#	sample	mass spectrometer	file size
1	Human serum	TOF	54 MB
2	Halobacterium salinarium	QTOF	333 MB
3	HeLa cell digest	LTQ Orbitrap	994 MB
4	Human urine	microTOF	2500 MB

Table 3.2: *Filtering for isotope pattern : the four LC-MS maps used in this experiment.*

Australia, and Professor Harald Mischak, Mosaique Diagnostics and Therapeutics, respectively.

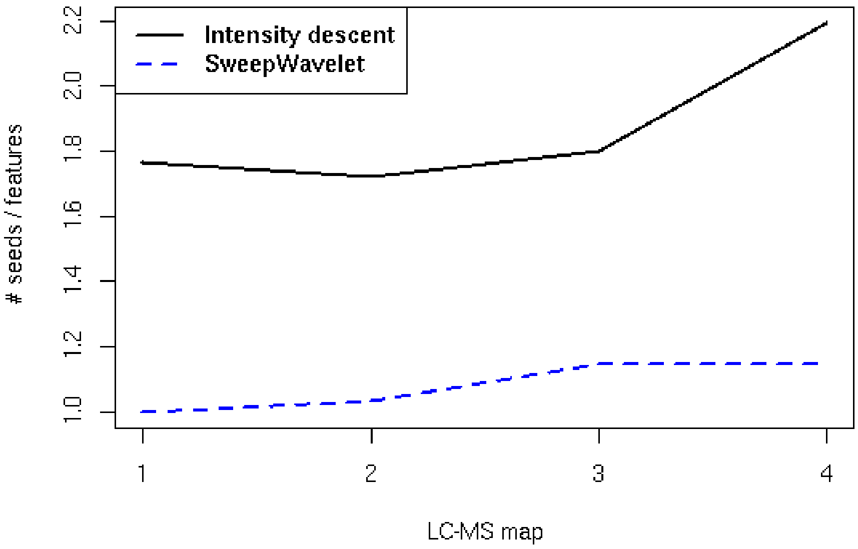


Figure 3.9: *Comparison of the number of seeds for* SweepWavelet *versus the intensity descent approach: the plot displays the ratio seeds / features (y-axis) versus the sample number. We fixed the number of features for all data sets.*

To assess the filtering performance of our approach, we counted the number of seeds that both algorithms needed to obtain the same number of peptide features. It is difficult to decide when a particular set of features is a sensible result. We can in most cases only make a judgment if we have some secondary information such as MS/MS spectra or a manual annotation. In this experiment, we decide to use the set of features computed by SweepWavelet as ground truth. We then performed a feature detection using our implementation of the intensity descent algorithm and stopped the algorithm if the same number of features was detected.

Fig. 3.9 shows the number of seeds required by both algorithms to obtain the same number of features on the four LC-MS maps. We see that the combination of isotope wavelet and averagine model clearly outperforms the intensity descent approach in terms of the number of seeds required. Especially for larger LC-MS maps, such as the urine sample recorded on a microTOF instrument, the intensity descent needs roughly twice the number of seeds to obtain the same number of features. In Chapter 4, we will show that intensity-descent algorithms also result in a higher number of false positives.

3.6.2 Charge Determination

There is so far no consensus in the literature on the best method to determine the charge of an isotope pattern (Kaur and O'Connor, 2006; Hoopmann *et al.*, 2007) but it seems that methods based on the Fourier transformation and matched filter have an advantage. In the experiment, we want to verify whether our combination of isotope wavelet and averagine model determines the charge of an isotope pattern correctly on data sets of different size and mass resolution.

To this end, we extracted six sets of peptide features from LC-MS maps recorded on different instruments (Grunert, 2008). We annotated each feature set with monoisotopic m/z, charge and intensity for each feature contained. After executing SweepWavelet on these data sets, we counted how many times our algorithm estimated the charge correctly.

#	sample	missed	% correct	annotated	Reference
1	TOF	9	98.03	61	(Mayr *et al.*, 2006)
2	TOF	1	85.00	22	(Mayr *et al.*, 2006)
3	TOF	1	85.71	15	(Mayr *et al.*, 2006)
4	QTOF	1	100.00	9	(Whitehead *et al.*, 2006)
5	LTQ Orbitrap	0	100.00	14	M. Robinson, WEHI
6	microTOF	7	100.00	23	H. Mischak, Mosaique

Table 3.3: *Charge determination : evaluation of* SweepWavelet*'s charge prediction on three data sets.*

Table 3.3 displays the results of this experiment. The table gives the number of annotated features, the percentage of annotated features for which the algorithm computes the charge state correctly and the number of features not detected (missed) by the algorithm. Our experiments reveal that our charge determination performs satisfactory. There is a slight drop in performance on the second and third data set which were recorded on an older TOF instrument. This is to be expected since charge determination on low resolution data is difficult even for a human expert.

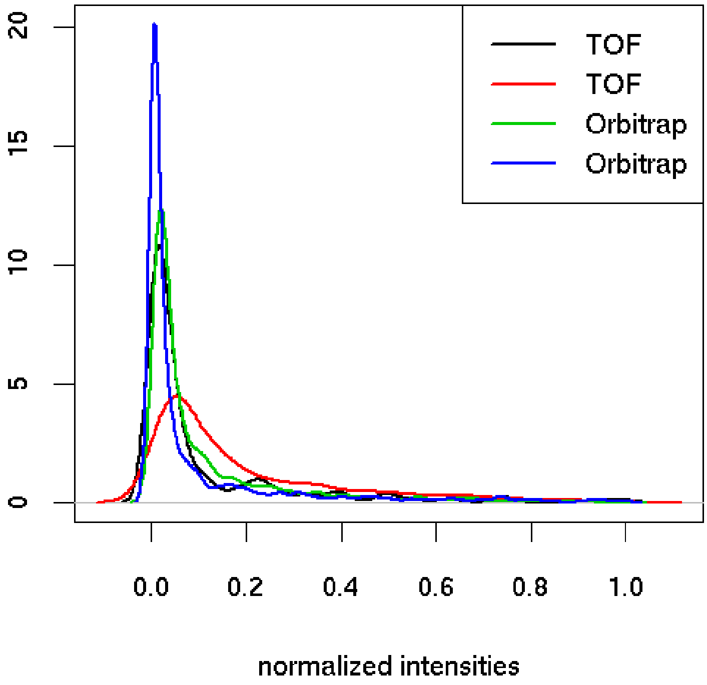


Figure 3.10: *Density plot of intensities extracted from four isotope pattern, recorded on TOF and Orbitrap instruments. We can see that the intensity distribution are not Gaussian, but have a pronounced tail to the right.*

We use Spearman's rank correlation coefficient to compute the goodness of fit between peptide signal model and the data. The rank correlation coefficient is a non-parametric measure of correlation and thus does not make any assumptions about the underlying distributions. This is in contrast to the standard Pearson correlation coefficient which assumes that the data follows a Gaussian distribution. This is certainly an advantage from a theoretical point of view. The question remains whether the intensities of isotope pattern do indeed require non-parametric approaches.

Figure 3.10 shows density plots of four isotope pattern recorded on TOF and Orbitrap instruments. The intensity distributions are clearly not Gaussian, but have a pronounced tail to the right. This makes sense since an isotope pattern consists of some elevated peaks which form the tail of the distribution but also a significant amount of background and lower intensities. Note that these isotope pattern were baseline-removed. Consequently, it seems to make sense to use a non-parametric measure of correlation for this task.

3.6.3 Classification of Peptide Expression Profiles

As we mentioned in the introduction to this section, we provide a detailed comparison of our approach for peptide quantification to other state-of-the-art algorithms follows in the next chapter. At this point, we want to give a proof of concept how quantitative mass spectrometry can be applied in a clinical setting.

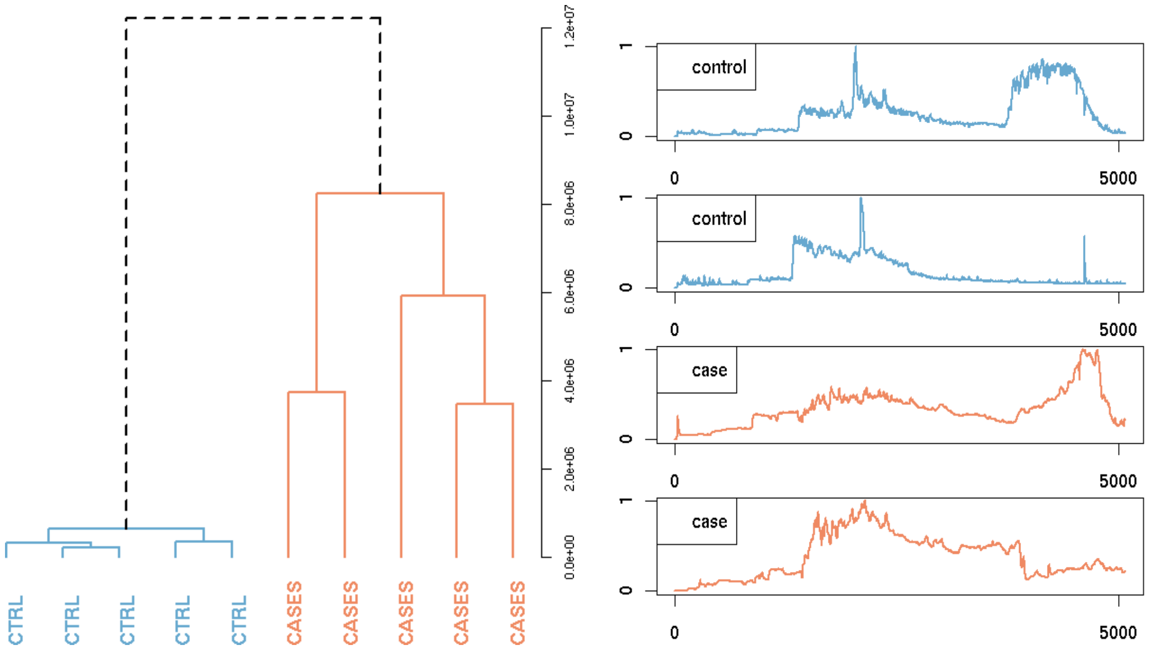
There are numerous studies in which an MS-based quantification of peptides or proteins was used for diagnostic purposes (Ge and Wong, 2008; Tibshirani *et al.*, 2004; Yu *et al.*, 2005). In most cases, this involves a classification problem: using data belonging to known groups (e.g. LC-MS maps obtained from healthy and diseased patients), we train a machine learning algorithm which is used to classify new samples and to distinguish healthy from diseased patients.

Building a reliable classifier for clinical purposes is a non-trivial task (Decramer *et al.*, 2008). The purpose of this section is to show that our algorithm SweepWavelet can be used to reliably classify clinical LC-MS samples based on their peptide expression profiles.

We used 10 maps obtained using Capillary Electrophoresis-Mass Spectrometry (CE-MS). This data was provided by Professor Harald Mischak (Mosaique Diagnostics and Therapeutics AG, Hannover). Capillary Electrophoresis (CE) is a separation method which separates charged molecules based on their mass-to-charge ratio and frictional forces (Kasicka, 2001). The capillary is filled with a buffer solution and has an electric field applied to it. The charged analytes move in this field: positively charged analytes are accelerated and negatively charged analytes are retarded. Accordingly, we refer to the time at which an analyte leaves the capillary as *migration time* (MT) instead of retention time. In a CE-MS setting, the capillary is directly coupled to an ESI mass spectrometer. CE offers a sensitive separation and is very cheap. The reason why it is not more frequently applied seems to be the poor reproducibility of the migration time (Kasicka, 2001).

Since the focus of this work is on LC-MS data, we will not further elaborate on the details of CE-MS but refer the interested reader to the literature (Chankvetadze, 1997; Decramer *et al.*, 2008; Kasicka, 2001). For our purposes, it suffices to know that the resulting data is essentially the same as an LC-MS map: we obtain a set of mass spectra, where each spectrum records a subset of the analytes leaving the capillary at a specific migration time. Five of the CE-MS maps were obtained from the urine of healthy volunteers and the remaining five from patients who underwent a renal allograft rejection.

To show that our algorithm SweepWavelet yields meaningful peptide expression values, we perform a proof-of-concept analysis of these samples. For each data set, we extracted the



(a) Dendrogram obtained by hierarchical clustering

(b) Total Ion Chromatograms

Figure 3.11: *Left: dendrogram obtained by a hierarchical clustering of the peptide expression profiles using complete linkage. Controls and cases are nicely grouped into separate subtrees. Right: total ion chromatograms for* 2 *control samples (top rows) and disease samples (bottom rows).*

peptide features and sorted them by intensity. We did not want to introduce any bias by an alignment algorithm and thus did not perform any dewarping of the retention times. Instead, we simply chose the 1000 most intense feature from each map and performed a hierarchical clustering using complete linkage on the feature intensities. We used the Euclidean distance as distance measure.

Figure 3.11 displays the results of this experiment. Figure 3.11(a) gives the dendrogram we obtained by hierarchical clustering. We can see that control and allograft rejection samples (cases) are well separated into different subtrees. It is also noteworthy that the subtree of control samples is denser, i.e., with shorter branch lengths than the subtree of the cases. The average euclidean distance between the feature intensities of the control samples is 7.63×10^5, whereas the average distance between the allograft rejection samples is 11.37×10^6.

This indicates that the disease samples are less homogeneous than the control samples which makes sense from the biological point of view. Figure 3.11(b) shows total ion chromatograms (TIC) of four LC-MS maps: two were obtained from control samples (top rows) and two from cases (bottom rows). Apart from the noisy region at the end of TIC, which we excluded during the feature detection, the TICs of the control samples exhibit a similar shape, whereas the cases are quite different and not many matching elution peaks can be identified.

This experiment is of course only a small case study. Nevertheless, other researchers demonstrated that that mass spectrometry-based diagnostics is feasible (Decramer *et al.*, 2008; Ge and Wong, 2008; Yu *et al.*, 2005) and we showed that our algorithm has the potential to be used for this task, too.

3.7 Summary

We presented SweepWavelet, an algorithm for the quantification of peptides from LC-MS data. Compared to other methods it has three novel aspects: first, it uses the isotope wavelet to rapidly filter spectra for isotope distributions. This mother wavelet also yields a first estimate of bounding box and charge. Second, we make efficient use of the redundant information in LC-MS maps and combine isotope pattern of the same peptide ion in adjacent spectra using a sweep-line approach. This approach helps us to improve our estimates of charge and monoisotopic m/z. Finally, the set of candidate isotope pattern in an LC-MS map is refined by comparing the extracted signals to a two-dimensional averagine model. This model is an improvement on previously applied methods since it models m/z peak shape as well as the elution curve of the peptide signal.

We applied SweepWavelet in three experiments: first, we investigated the filtering performance of the isotope wavelet. We compared our algorithm to an implementation of an intensity descent which selects the seeds not by shape but by intensity. We counted the number of seeds that both algorithms need to find the same number of features on LC-MS maps of different sizes. SweepWavelet always requires less seeds to obtain the same number of features. This is important since more seeds translate into more average model fitting attempts. This will in turn translate into more false positives since some signals will incorrectly be declared as features due to the high number of attempts.

Second, we tested the charge detection performance of our algorithm and found it to be satisfactory. Accurate charge determination of features is important in different scenarios such as estimation of the corresponding protein abundance or sequencing of the MS/MS spectrum associated with this feature.

Finally, we applied SweepWavelet in a classification task: the aim was to performed an unsupervised clustering of CE-MS maps obtained from healthy volunteers and probands who experienced a renal allograft rejection. Hierarchical clustering performed on the feature intensities resulting in an excellent separation of cases and controls. This result agrees with other publications which show that mass spectrometry-based quantification has clinical potential (Decramer *et al.*, 2008; Ge and Wong, 2008; Yu *et al.*, 2005).

The next chapter contains detailed benchmarks of our algorithms and a comparison to other state-of-the-art approaches on real and simulated data sets.

Chapter 4

Benchmarking Quantification Algorithms

Synopsis: *This chapter presents a comparison of our algorithm,* SweepWavelet*, and competing methods. We use real and simulated LC-MS maps to perform benchmarking studies and give insights into the strengths and weaknesses of current approaches. Due to the inherent complexity of LC-MS data sets, these benchmarks are a difficult task. To our knowledge, this is the first attempt to compare the relative strengths and weaknesses of LC-MS feature detection algorithms so far.*

4.1 Introduction

In this chapter, we compare the performance of SweepWavelet, our algorithm for peptide quantification to existing ones. For practical purposes, a good feature detection and quantification algorithm should accurately determine the positions of all features in rt and m/z and give a good estimate of the abundances. Parts of the results presented in this chapter are already published (Schulz-Trieglaff *et al.*, 2007, 2008a,b).

Even if peptide feature detection and quantification is a fundamental step in a proteomics data analysis workflow, so far no comparison of feature detection algorithms has been made. We believe that such a comparison would be beneficial since it seems that most algorithms are tailored for specific mass spectrometers and their data. Some are very fast but inaccurate, other only detect features but perform no reasonable quantification etc. Knowledge about the specific advantages and disadvantages of current algorithms would allow each user to find the optimal algorithm for his or her data and instrument settings.

Nevertheless, comparing LC-MS feature detection algorithms is not a straightforward task. We can record mixtures of peptides in different concentrations and investigate if the algorithm recovers the differences in abundance. The problem is that even for simple mixtures, it is difficult to determine in advance which theoretically possible peptides will appear where in the LC-MS map and consequently, a manual annotation is required. On the other hand, such an annotation is tedious for all but the most simple data sets. But simple mixtures usually do not pose a challenge for an algorithm. For more complex samples, the ground truth e.g. which peptides will actually appear and in which relative abundances is even more difficult to determine. In this case, one possibility is a semi-automatic annotation using MS/MS sequencing. But this is also not a fail-safe procedure: MS/MS sequencing will only identify a subset of peptides correctly and thus cannot be relied on for an annotation of complex samples. To give an example, our algorithm SweepWavelet detects more than 5000 high-confidence features in an LC-MS map of a tryptic digest of human HeLa cells. 1200 could be confidently identified using the InSpect software (Tanner *et al.*, 2008). Undeniably, it is difficult to make a statement about the remaining signals.

All theses reasons make it difficult to assess the results of a quantification algorithm. To remedy this, we follow two different routes: we compare our algorithm to existing ones on two real data sets of intermediate complexity. We even have a partial annotation for the first data set, several LC-MS recordings of human plasma. Furthermore, we use simulated data to benchmark the algorithms on a more fine-grained level.

We compare our algorithm, SweepWavelet (Schulz-Trieglaff *et al.*, 2007), to the algorithms Superhirn (Mueller *et al.*, 2007), msInspect (Bellew *et al.*, 2006), SpecArray (Li *et al.*, 2005), MZmine (Katajamaa *et al.*, 2006) and Decon2LS[1] that were all introduced in the previous chapter. We did not consider Hardkloer (Katajamaa *et al.*, 2006), AID-MS (Chen *et al.*, 2006), lcms2d (Chen *et al.*, 2006) and MapQuant (Leptos *et al.*, 2006). Hardkloer and Decon2LS are both re-implementations of the THRASH algorithm and we decided to consider only one of them. The remaining tools are tuned for specific instruments and their data formats. lcms2d crashed on most of our data whereas AID-MS and MapQuant use a proprietary data format ('XCaliburRaw' by Thermo Finnigan) and were not able to read LC-MS maps in any other format.

[1] http://ncrr.pnl.gov/software/

4.2 Benchmarks

We believe that good quantification algorithms should be able to deliver reasonable results on data from different instruments and not be tuned for a particular LC-MS setup. Consequently, we tested the aforementioned algorithms on data sets from different instruments and mass resolutions. First, we use real data to assess the quantification performance. These are LC-MS runs from a Linear Ion Trap with intermediate resolutions and a high-resolution data set from a microTOF instrument. Second, we perform a more fine-grained comparison of the different algorithms on simulated data. We simulate different resolutions and chromatographic conditions and assess the accuracy of feature detection.

4.2.1 Absolute Quantification of Myoglobin in Human Serum

Myoglobin is a globular protein which is present in the cytosol of the cardiac and skeletal muscle. Its main task is the storage and transport of oxygen in muscle tissue. Myoglobin is released into the blood shortly after a myocardial infarct and is therefore one of the earliest known biochemical marker for myocardial necrosis and infarction, i.e., after an heart attack (Kagen *et al.*, 1975). The aim of this project was to develop a reference method for the automated quantification of myoglobin in human serum.

We used data from a mass spectrometry-based myoglobin quantification study (Mayr *et al.*, 2006). To obtain a ground truth for the evaluation of our algorithms, we removed all naturally occurring myoglobin from a set of human serum samples. The myoglobin-depleted human serum and a stock solution of human myoglobin was provided by IRMM (Institute for Reference Materials and Measurements, European Commission, Geel, Belgium). Highly abundant serum proteins can render an accurate quantification impossible since they might overlap with or even suppress the signals of less abundant proteins such as myoglobin. To this end, we removed highly abundant serum proteins from the sample by anion-exchange chromatography. Subsequently, we added a known amount of human myoglobin to aliquots of the sample. This yields our ground truth, e.g. the target value for quantification. Using this ground truth value, we were able to compare the results of different algorithms.

Note that we are aiming at an absolute quantification, as compared to a relative quantification which only yields expression ratios. To achieve this, we added further amounts of myoglobin to aliquots of the serum samples. As an example, the first data set consists of 8 groups, each group consists of 4 aliquots. Each group received a different amount of myoglobin, ranging from 0.0 to 2.836 ng/μl. Finally, we added a fixed amount of horse myoglobin as an internal standard. This helps us to account for variability in overall serum composition, sample preparation and measurement of the analyte (Mayr *et al.*, 2006). The mixture was enzymatically digested using trypsin.

A human expert performed a manual annotation of this data set. This means that we know the positions in m/z and rt of the myoglobin peptides that we need for quantification. We determined the ground truth concentration of myoglobin by performing a linear regression on the signal intensities of selected myoglobin peptides. It turned out that the ratio of the signal intensity of the eleventh myoglobin peptide and the corresponding (homologous) horse peptide yielded the best results and we used this approach for all algorithms. Figure 4.1 illustrates the absolute quantification. A linear regression is performed on the ratio of the feature intensities of the two myoglobin peptides (human and horse). The estimated concentration is the absolute value of the x-axis intercept of the regression line. We conducted this experiment two times, resulting in two groups of LC-MS maps consisting of 32 maps each.

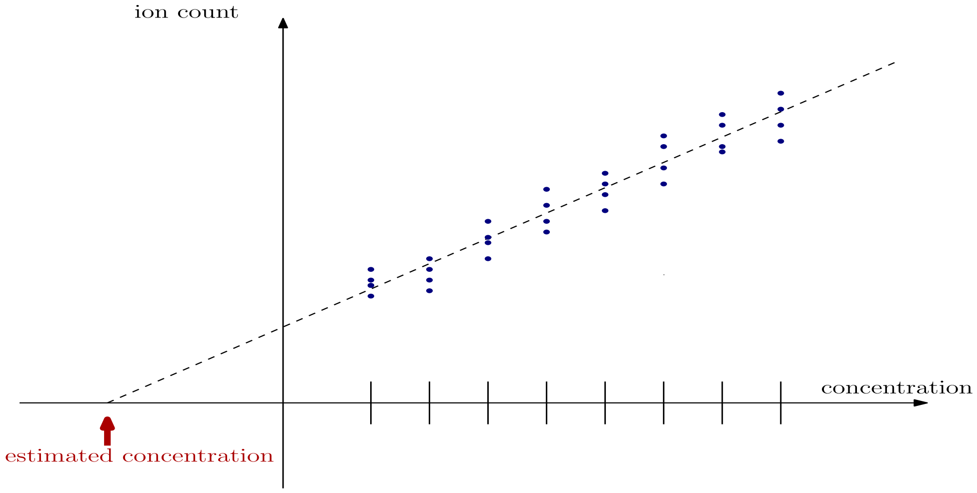


Figure 4.1: *Illustration of the absolute quantification using linear regression. For each of the eight spike-in concentrations, we measured four replicate LC-MS maps and determined the ratio of the feature ion counts for the eleventh myoglobin and the eleventh horse myoglobin peptide (blue dots) in each map. The estimated myoglobin concentration is the absolute value of the x-axis intercept of the regression line.*

The LC-MS maps were recorded using a reversed-phase HPLC column coupled to a Quadrupole ion trap mass spectrometer (model Esquire HCT, Bruker Daltonics, Germany). This instrument yields mass spectra with low to intermediate mass resolution and accuracy. For some few signals, well-resolved isotope patterns are discernible, but in the majority of the cases the signals are poorly resolved and have a more or less unimodular and skewed shape. Table 4.2.1 summarizes the results of all algorithms. For comparison we also give the results of a manual quantification: rows labelled with 'Manual' indicate the results obtained by a human expert (Mayr *et al.*, 2006). The manual quantification was performed using Bruker's esquireData software and Microsoft Excel 2000. Peak areas were estimated from extracted ion chromatograms smoothed by a Gaussian filter. Among the algorithms we tested, only MZmine and SweepWavelet perform a reasonable quantification on both data sets. The quantification results of SweepWavelet differ from the results published in Schulz-Trieglaff *et al.* (2008a) since we used an improved version of the averagine matching module. Decon2LS finds several myoglobin peptides but their abundance estimate is not even close to the true value. SpecArray performs only well on the first data set. Superhirn and SpecArray fail to find any myoglobin peptides in the second data set and where thus not able to perform any quantification at all. Both tools are tailored for high-resolution data and fail to recognize the weaker myoglobin features as meaningful signals.

The myoglobin experiment represents our first test case for feature detection algorithms. Even if the samples were filtered for highly-abundant serum proteins, they were still of surprising complexity and several hundreds of peptide ions can be discerned by eye. This experiment nicely shows that newer algorithms, such as Superhirn and SpecArray, are tuned for higher resolution data sets recorded on newer mass spectrometers such as Orbitraps or FT instruments.

Nevertheless, this specific absolute quantification task does not represent an ideal benchmark for a quantification algorithm. From a computational point of view, the task is to detect and quantify two signals (the two myoglobin peptides on which quantification was performed) as accurately as possible. But we cannot say much about the remaining features extracted. Furthermore, the ion trap instrument used in this study is not state-of-the-art anymore. That is why we proceed with a more detailed evaluation in the next sections.

	Estimate [ng/μl]	95% conf. interval [ng/μl]	Rel. deviation [%]	r^2
Myo # 1	**True concentration** [ng/μl] *0.463*			
Manual	0.382	[0.315;0.454]	−17.494	0.991
Decon2LS	−3.094	[0.933;41.571]	−768.250	0.076
msInspect	0.386	[0.290;0.491]	−16.630	0.957
MZmine	0.448	[0.369;0.547]	−3.239	0.960
SpecArray	0.405	[0.271;0.586]	−12.569	0.938
Superhirn	–	–	–	–
SweepWavelet	0.462	[0.340;0.599]	−**0.215**	**0.951**
Myo # 2	**True concentration** [ng/μl] *0.463*			
Manual	0.382	[0.315;0.454]	−17.494	0.991
Decon2LS	−4.758	[−1.009;8.936]	−1127.645	0.0223
msInspect	0.170	[1.469;4.988]	$-63,282$	0.957
MZmine	0.448	[0.377;0.525]	**−3.239**	**0.970**
SpecArray	–	–	–	–
Superhirn	–	–	–	–
SweepWavelet	0.488	[0.297;0.728]	$5,339$	0.825

Table 4.1: *Results from the myoglobin quantification study. We give results for all algorithms and the manual quantification. We give the estimated abundance, confidence interval and Pearson correlation as estimated from the linear regression. The relative deviation (column* 4*) from the true value is given by* $100 \times (x_{measured} - x_{true})/x_{true}$.

4.2.2 Relative Quantification of Standard Peptide Mixes

The aim of this experiment was to evaluate the performance of our algorithm in a more complex task: the quantification of peptides from a mixture of known composition. Eight proteins, purchased from Sigma (St. Louis, MO) and Fluka (Buchs, Switzerland), were digested using trypsin (Promega, Madison, WI) using published protocols (Schley *et al.*, 2006). The mixture contained the following proteins in concentrations between $0.5 - 2.5$ pmol/μl: β-casein (bovine milk), conalbumin (chicken egg white), myelin basic protein (bovine), hemoglobin (human), leptin (human), creatine phosphokinase (rabbit muscle), α1-acid-glycoprotein (human plasma) and albumin (bovine serum). The resulting peptide mixture was then separated using capillary IP-RP-HPLC and subsequently analyzed by Electrospray Ionization Mass Spectrometry (ESI-MS) as described in Schley *et al.* (2006); Toll *et al.* (2005). The separation was carried out in a capillary/nano HPLC system (Model Ultimate 3000, Dionex Benelux, The Netherlands) using a 50 × 0.2 mm monolithic poly-(styrene/divinylbenzene) column (Dionex Benelux) and a gradient of 0-40% acetonitrile in 0.05% aqueous trifluoroacetic acid for 60 min at 55°C. The digest was analyzed in triplicate at a flow rate of 2μl/min. The injection volume was 1μl. On-line ESI-MS detection was carried out with a time-of-flight mass spectrometer (model microTOF, Bruker Daltonics) in positive ion mode. We obtained three replicate measurements with the same concentration of each protein spiked-in.

In contrast to the myoglobin data set, we do not know which peptides will appear in these LC-MS maps since we do not have a manual annotation. But we can make some less stringent assumptions. First, a true positive peptide feature should have a mass closely matching the mass of a theoretical peptide. If an algorithm computes many features with masses strongly deviating from theoretical peptide masses, this indicates a less accurate feature detection process. The *mass deviance* is defined as the minimum distance between the mass of a feature and the closest peptide mass obtained from a theoretical protein digest. Furthermore, since we have three replicates of the peptide mixture, we can also compare the feature intensities which should be approximately equal across the replicates. Finally, we can make some statements about the number of features that we would expect to find. A theoretical digest of the proteins used for this experiments yields 830 peptides allowing a maximum of two cleavage sites. Since we used an ESI mass spectrometer, we expect to obtain more than one charge state variant per peptide, depending on size and composition. Of course, not all of these peptides will appear in the LC-MS map. Our experiments have shown that roughly 30-50% of all theoretically possible peptides will actually appear in the LC-MS map. If we assume that each peptide forms two charge variants on average during ionization and that 50% of these ions will appear in the LC-MS map, we can conclude that ≈ 800 peptide signals should be detected. Interestingly, the number of features detected by each algorithms differ widely. To give an example, MZmine detects 621 on the first LC-MS map, whereas msInspect detects 12005. This gives us already an indication that the algorithms differ widely in their selectivity.

	mean mass dev.	stdev. mass dev.	# features
Decon2LS	30.13/34.68/31.82	170.98/200.63/205.12	5602/5934/5480
msInspect	380.03/382.54/335.31	1190.12/1212.79/1104.98	12005/11624/11792
MZmine	10.25/9.57/10.17	30.83/32.22/35.38	2711/2741/2685
SpecArray	15.81/18.60/14.90	49.10/65.58/60.25	1334/1239/1269
Superhirn	29.24/28.96/26.62	69.73/69.42/67.06	3571/3652/3565
SweepWavelet	**9.06/9.46/9.12**	**27.99/28.53/28.95**	1544/1601/1266

Table 4.2: *Comparison of feature detection algorithms on the standard peptide mixture (Schley* et al.*, 2006). We give the results in terms of the mass deviance, the minimum distance of a feature mass to a theoretical peptide mass. The third column gives the number of features detected for each of the three replicates.*

We executed all feature detection algorithms on the three replicate LC-MS recordings of the peptide mixture. Table 4.2 displays the results in terms of mean and standard deviation of the mass deviance. The third column gives the number of features detected. We can see that several algorithms compute features with a very large mass deviance. Some of these features can be peptides that we did not consider, such as trypsin peptides resulting from a self-digest of the enzyme. Since we compute the mass deviance using the decharged mass instead of m/z, high values for the mass deviance can also result from incorrect charge assignments. Nevertheless, most of the features with a high mass deviance are likely to be false positives.

	s1 vs s2	s1 vs s3	s2 vs s3
Decon2LS	0.89 (2725/48.64)	0.86 (2433/41.00)	0.93 (2821/51.47)
msInspect	**0.99** (12009/43.50)	0.95 (11628/42.19)	0.95 (11796/43.43)
MZmine	0.99 (625/23.00)	0.95 (611/22.29)	**0.97** (640/23.83)
SpecArray	0.99 (1338/82.43)	0.96 (1243/74.56)	0.96 (1273/81.49)
Superhirn	0.96 (2150/0.60)	0.91 (2064/0.56)	0.94 (2127/0.60)
SweepWavelet	0.98 (981/70.12)	**0.98** (762/51.97)	0.96 (866/75.17)

Table 4.3: *The Pearson correlation coefficient between each pair of replicates. The numbers in brackets give the absolute number of features and the percentage remaining after alignment.*

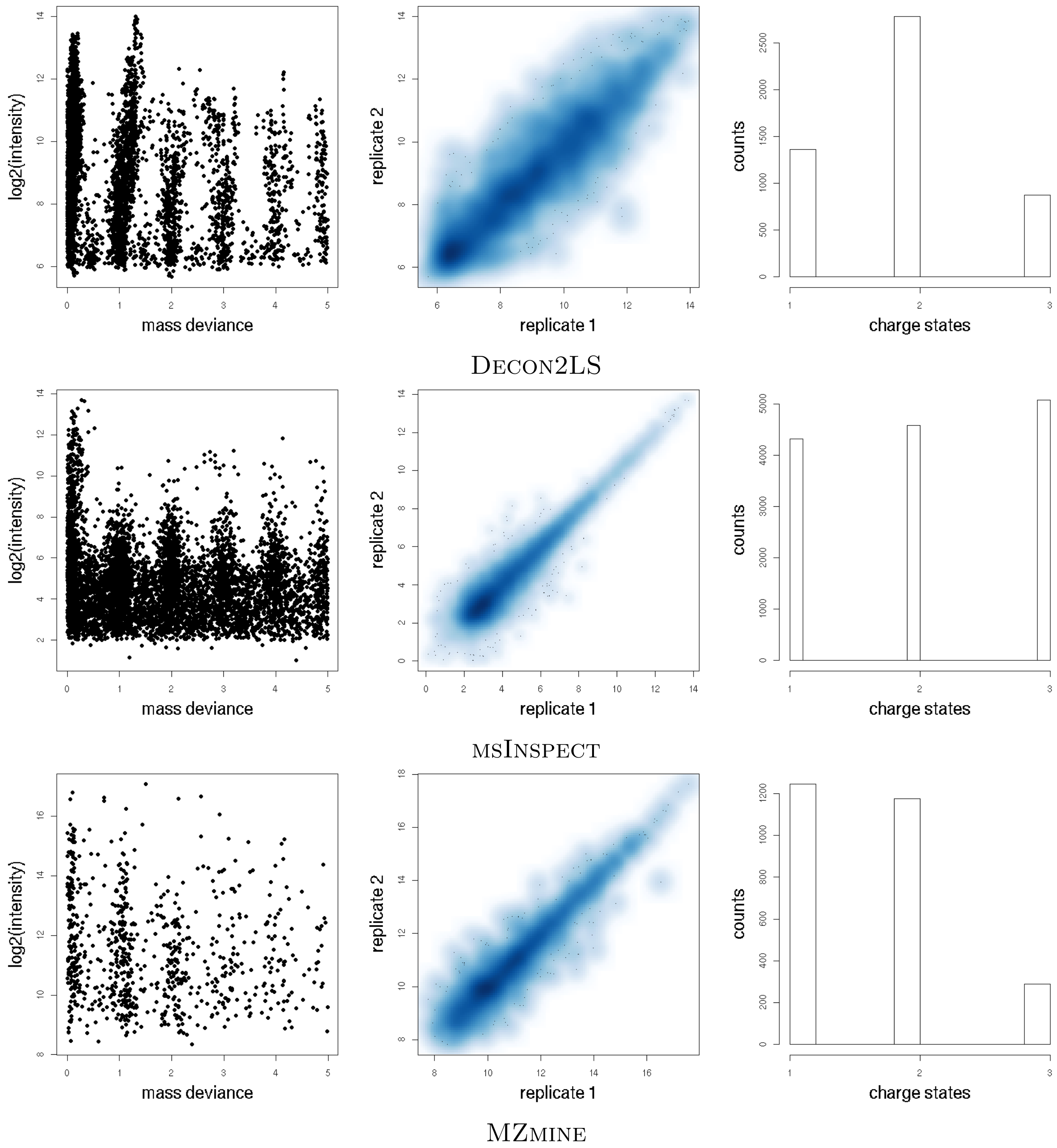
In sum, we see that msInspect and Decon2LS are relatively inaccurate. Especially msInspect computes huge numbers of false positives. Since msInspect is an integrated package combining feature detection and alignment, we believe that this is part of its strategy. False positives will be features appearing at inconsistent positions after alignment and can thus be filtered out during later steps of the data analysis. Nevertheless, this will unnecessarily exacerbate the alignment and we will show below that alignment does not reduce the number of false positives by a large amount.

Finally, the algorithms SpecArray and Superhirn yield comparable results in terms of the mass deviance but lag behind MZmine and SweepWavelet which show a similar performance. SweepWavelet stands out as the best algorithm on all data sets in terms of mean and standard deviation of the mass deviance.

To compare the quantification performance, we computed the pairwise correlation coefficients between the feature intensities of the replicate measurements. We obtained an alignment using the algorithm implemented in OpenMS (Lange *et al.*, 2007). We used the standard parameters for each alignment. Table 4.3 shows the results. All algorithms yield similar results, only Decon2LS lags significantly behind. The numbers of brackets in each column gives the absolute number of features remaining after alignment and the percentage of features remaining, respectively. The alignment drastically reduces the number of features: in the case of Decon2LS, less than 3000 (50%) remain. Still, the average mass deviance of the remaining features decreases from 30.13 to 26.54 only. This number is still high. This might be an indication that alignment is not successful in removing all false positive features and that it might be more important to avoid false positives in the first place.

Figure 4.2 and 4.3 summarize the results on the peptide mixture. For each algorithm we examined, the first column gives a plot of the mass deviance versus the log-scaled intensity on the first replicate. We give only values in the range from 0 to 5 since features with a higher deviance are extreme outliers. Nevertheless, some of the algorithms, especially msInspect and Decon2LS, detect a lot of features with mass deviances outside this range. The plots reveal that features with high mass deviance tend to be in the lower intensity region. This is not surprising since one challenge is to find true features of low intensity just above the noise. Simply not considering signals with low intensity is not an option since they might be of high significance e.g. as a biomarker. Furthermore, we can see that most feature masses seem to form clusters at regular intervals. This is a phenomenon that has been observed in other studies as well (Du and Angeletti, 2006; Piening *et al.*, 2006; Wolski *et al.*, 2006). It is due to the fact that peptides are a relatively homogeneous class of molecules and that the 21 proteinogenic amino acids result only in a limited number of possible peptide masses.

The second column in Fig. 4.2 and 4.3 gives for each tool a plot of the feature intensities of the first replicate versus the second replicate LC-MS map. The results for the remaining replicate maps are similar (see Table 4.3). We can see that most tools achieve a good quantitation but some outlier at lower intensities. The third column gives the distribution of charge states detected by each algorithm. For a sample of this composition, we would expect that most peptide ions occur in charge states 1 to 3. Most tools assign the majority of peptides to these charge states. All higher charge states are likely to be incorrect. Note that MZmine only considers charge states from 1 to 3 by default and does not implement charge detection routines for higher charge states. That is why we restricted to charge state range for all tools to $[1, 3]$.

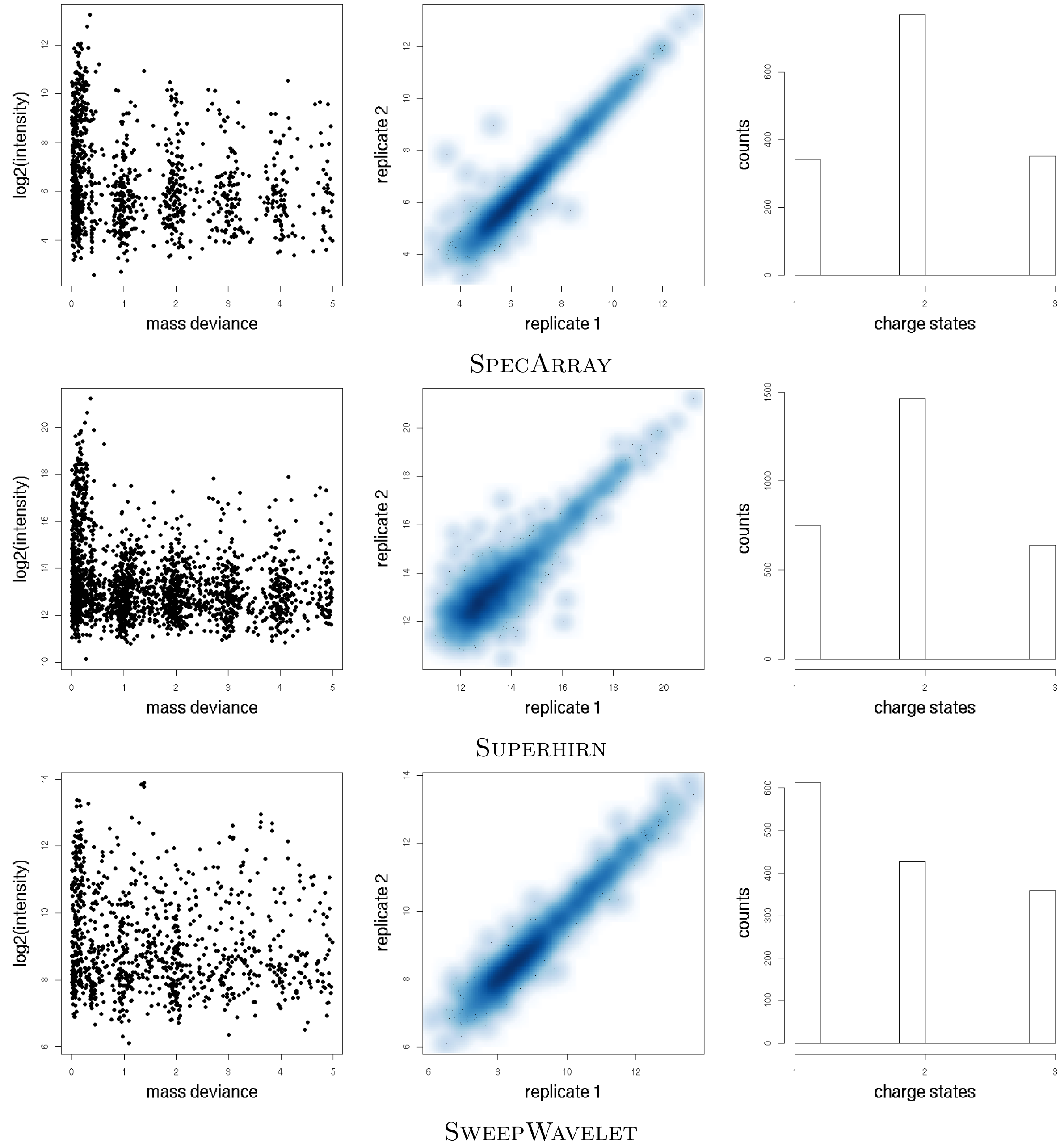


Decon2LS

msInspect

MZmine

Figure 4.2: *Comparison of mass deviance, reproducibility of intensities and charge detection of the tools* Decon2LS, msInspect *and* MZmine. *For the mass deviance, we give only the range from* 0 *to* 5 *since all other features are extreme outlier.* MZmine *only considers charge states from* 1 *to* 3 *by default and does not implement charge detection routines for higher charge states. That is why we restricted the charge state range for all other tools to* $[1, 3]$*, too.*



SpecArray

Superhirn

SweepWavelet

Figure 4.3: *Comparison of mass deviance, reproducibility of intensities and charge detection of the tools* SpecArray*,* Superhirn *and* SweepWavelet*. For the mass deviance, we give only the range from* 0 *to* 5 *since all other features are extreme outlier.* MZmine *only considers charge states from* 1 *to* 3 *by default and does not implement charge detection routines for higher charge states. That is why we restricted the charge state range for all tools to* $[1, 3]$*, too.*

4.2.3 Simulation Benchmarks

At this point, we already have an idea of the strengths and weaknesses of current feature detection algorithms. Decon2LS and msInspect are prone to compute many false positives whereas SpecArray and Superhirn are highly specific but do not perform well on low resolution data such as the myoglobin maps. We have to be aware that the peptide mixture and the myoglobin data are on opposite ends of the range of mass spectrometry instruments that are currently available. Whereas the myoglobin data was recorded on a low-resolution instrument, the peptide mixture was recorded on a modern instrument with high mass resolution and accuracy. One might argue that older instruments will be replaced at some point. But due to monetary restrictions, they are still the workhorses of many labs and will be for the immediate future. On the other hand, there are quite a few instruments which produce data that lies in between the two extremes that we examined in the previous to sections. In this section, we introduce our simulation model for LC-MS data which is implemented in the open source software LC-MSsim. We will use this software to create simulated LC-MS maps and perform benchmarking experiments using these maps. The following sections are published in Schulz-Trieglaff *et al.* (2008c)

We would like to emphasize that our aim was not to create a detailed physical model of mass spectra generation as, for instance, attempted in Coombes *et al.* (2005). But we want to simulate data that is reasonably close to reality and provides a fair testing ground for data analysis methods. The idea of simulating ESI mass spectra to assess the performance of MS feature detection algorithms was pioneered by Wong and Downard (2005). They simulate ESI spectra as mass lists derived from theoretical digests of protein sequences with normalized intensities without prediction of ion intensities, retention times or simulation of isotope pattern. They also restrict their comparison to their own algorithm which implements a very specific task, the detection of protein-ligands and other macromolecular complexes in mass spectra. Of course, the applications of LC-MSsim are not restricted to feature detection benchmarks. The next obvious step would be to compare alignment algorithms, but even the comparison of a full quantification workflow is an interesting scenario.

To our knowledge, LC-MSsim is the first software that models the whole LC-MS data acquisition process and delivers an output (the simulated LC-MS map and the list of peptides and contaminants with m/z and retention time) that can directly be used for the assessment of proteomics algorithms. There are, of course, some programs that simulate individual parts of the LC-MS data acquisition process, such as the estimation of isotope peak patterns[23], the prediction of peptide retention times (Meek, 1980; Petritis *et al.*, 2006; Krokhin, 2006; Klammer *et al.*, 2007; Pfeifer *et al.*, 2007) or detectability (Mallick *et al.*, 2007; Tang *et al.*, 2006). But these tools are written in different programming languages, they write different output and cannot easily be combined. However, to simulate a full LC-MS run, it is clearly desirable to have all these tools combined in a single application. We give implementation details and installation instruction of LC-MSsim in the Appendix.

An artificial LC-MS data set is generated by the following steps:

1. protein digestion,
2. prediction of peptide detectability and retention time,
3. determination of charge states,

[2] ProteinProspector: http://prospector.ucsf.edu/

[3] IsotopIdent: http://education.expasy.org/student_projects/isotopident/

4. modeling of isotope and elution profiles
5. and addition of noise (m/z, rt error and shot noise) and contaminants.

Key parameters that influence the outcome of the simulation are

- the minimum accepted peptide detectability (influences the number of theoretical peptides appearing in the LC-MS map),
- mass accuracy and resolution,
- the Full-Width-At-Half-Maximum (FWHM) of the peptide peaks,
- percentage of non-peptide contaminants,
- length of the gradient and m/z range of the mass spectrometer
- the noise level.

In the following sections, we give an overview of all simulation steps and explain their parameters in more detail.

Protein Digestion

The user can supply a list of protein sequences in a FASTA file and define their relative abundances in the sequence header. If no abundance is given, we assume that each protein, and thus its peptides, will appear in equal abundances in the mass spectra (apart from effects such as ion suppression etc.). LC-MSSIM supports only tryptic digests in this version, but new proteases can be added easily by extending the corresponding OpenMS classes by new regular expressions. We can also simulate missed cleavages and self-digestion of the protease.

Prediction of Peptide Detectability and Retention Time

After the enzymatic digest of all protein sequences, we need to determine the retention time of each peptide. Pfeifer *et al.* (2007) recently introduced the *paired oligo-border kernel* (*POBK*) for machine learning problems in computational proteomics. Support vector regression (Schölkopf *et al.*, 2000) using this kernel function yields very accurate retention time predictions while requiring only a small number of training samples. We use the POBK for retention time prediction in our simulator. We trained the SVM on the test set of Petritis *et al.* (2006) and determined the best parameters using nested cross-validation. The data set consists of 1304 peptide identifications of capillary reversed-phase liquid chromatography runs.

It was previously shown that not all peptides of a digested protein actually appear in the LC-MS sample (Mallick *et al.*, 2007; Sanders *et al.*, 2007; Tang *et al.*, 2006). There are numerous reasons for that. Some peptides will only poorly ionize in the electrospray and others are simply not soluble, just to give an example. To account for this fact, we determine the likelihood of detectability of each peptide using a support vector machine (Vapnik, 1995) and the POBK as a kernel function. We performed a nested cross-validation on balanced samples of the MUDPIT-ESI data set of (Mallick *et al.*, 2007). This means that we selected all n positive examples and chose n negative examples randomly out of all negative examples. The whole process was repeated ten times to minimize random effects. Mallick *et al.* (2007) evaluated their method in terms of 1 - positive predictive value (PPV) against coverage. For the MUDPIT-ESI data set they got a coverage of about 0.65 at $1 - \text{PPV} = 0.1$ and a coverage of about 0.75 at $1 - \text{PPV} = 0.15$.

In our evaluation (data not shown) we achieve a coverage of about 0.65 at $1 - \text{PPV} = 0.11$ and a coverage of about 0.8 at $1 - \text{PPV} = 0.15$. This means that our method performs comparable to the methods of Mallick *et al.* (Mallick *et al.*, 2007) although requiring just a fraction of the negative data set for training (about 1000 instead of $25,000$), which drastically decreases the time needed for training the classifier. We used the probability estimates (Wu *et al.*, 2003) of the libsvm library (Chang and Lin, 2001) to compute the likelihood of a peptide to appear in the LC-MS spectra.

Determination of Charge States

After protein digestion, peptide detectability and retention time prediction, we need to determine the relative abundances of the ions created for each peptide. LC-MSSIM models an Electrospray Ionization (ESI) mass spectrometer. ESI ionizes peptides and other sample compounds by applying a strong electric field to the sample. This field induces a charge accumulation at the liquid surface which will form highly charged droplets. As a result, we expect to see one to four ions per peptide, but charge states two or three are the most common. There are several chemical models describing the charge distribution for molecules after ESI and numerous factors influence this distribution such as the pH, sample composition and conformation of the peptide (Iavarone *et al.*, 2000; Konermann, 2007). However, our experiments have shown that a simple model gives a good approximation of real data.

For this reason, we decided to stick with a straightforward model of an ESI mass spectrometer in positive ion mode. We follow an approach by Schnier *et al.* (1995) and assume that each basic amino acid in a peptide can receive at most one charge unit (proton). Consequently, most tryptic peptides have a maximum charge state of $2 - 3$ which matches observations of real data. We determine the relative abundances of each charge state by sampling from a binomial distribution. As a result, low charge states are much more likely to occur than higher ones.

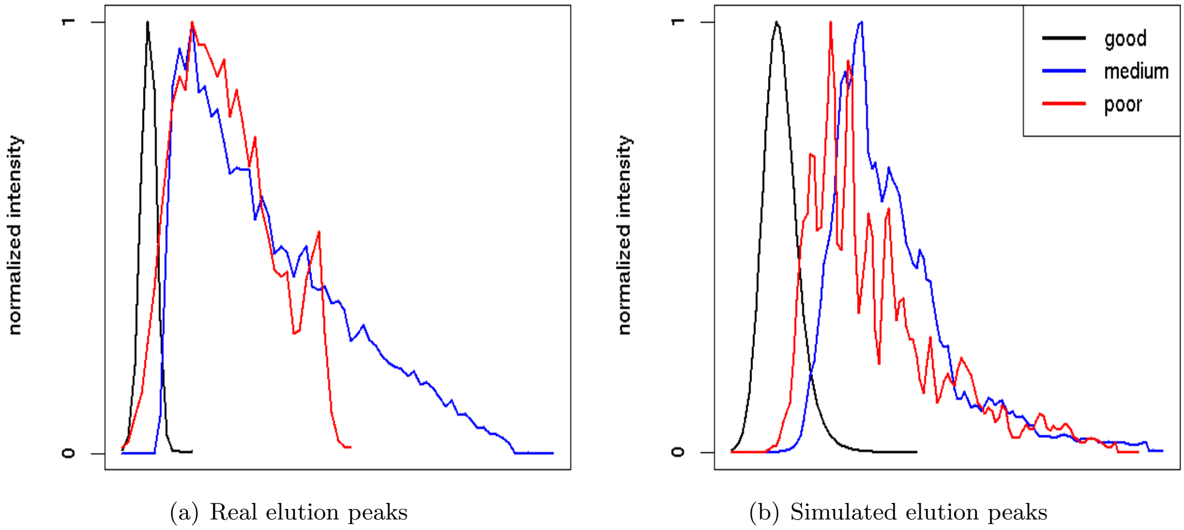
Modeling of Isotope and Elution Profiles

The position of a peptide ion signal in the LC-MS map is determined by three parameters: monoisotopic mass, charge and retention time. We calculate the mass from the amino acid sequence, charge is given by our binomial charge distribution model and the retention time predicted by the SVM.

Usually, a peptide ion gives rise to several peaks in a mass spectrum due to the fact that some of its atoms will occur in heavier isotope states. Given the sequence of the peptide, we calculate its monoisotopic mass from its empirical formula. The relative heights of the isotope peaks are calculated using a simple but fast algorithm (Kubinyi, 1991). This algorithm gives us the relative intensities of the isotope peaks. We model the peak shape using a Gaussian distribution. The user can choose the peak width in terms of the Full-Width-At-Half-Maximum (FWHM). The FWHM of a peak in a mass spectrum is given by the difference of the m/z values at which the ion count equals half of the maximum ion count of this peak. Note that we assume the peak shape to be Gaussian. The FWHM of a Gaussian function is given by $2\sqrt{2\ln 2}\sigma$, where σ is the standard deviation of the Gaussian.

Whereas the shape of peaks in the m/z dimension is relatively stable during one experiment, the peak shape in retention time might vary considerably, but has often a Gaussian-like shape. To account for this fact, we model the elution profile of a peptide signal using different chromatographic functions: a simple Gaussian distribution and an exponentially modified Gaussian distribution (EMG) (Grushka, 1972). Whereas the Gaussian function represents a perfect chromatographic condition, the EMG can model different distortions of the elution peak. Its exponential component introduces tailing and fronting effects. We described more details of the EMG function in Chap. 3.

Furthermore, we add uniformly distributed noise to single sampling points of the EMG but smooth the noisy elution profile afterwards to obtain ragged chromatographic peaks. This allows us to model more realistic chromatographic peaks as an elution profile in a real LC-MS run is never entirely smooth. On the other hand, this introduces several additional parameters into the simulation. To make the software more user-friendly, we supply a set of pre-defined parameter sets for the EMG, entitled *poor*,*medium* and *good* chromatographic conditions. A choice of good conditions leads to almost perfect Gaussian-shaped peaks, like they will almost never appear in a real experiment. Accordingly, medium and poor conditions lead to far more noisy elution peaks, including tailing and fronting effects. The corresponding peptide signals will be more difficult to trace across all their spectra since most algorithms have problems to trace features with very rough elution peaks. Figure 4.4 shows three elution peaks from a reversed-phase column and as well as three simulated elution profiles, one for each parameter set. As we can see, the simulated peaks are close to real elution peaks and cover a sufficiently broad range of chromatographic conditions.



(a) Real elution peaks

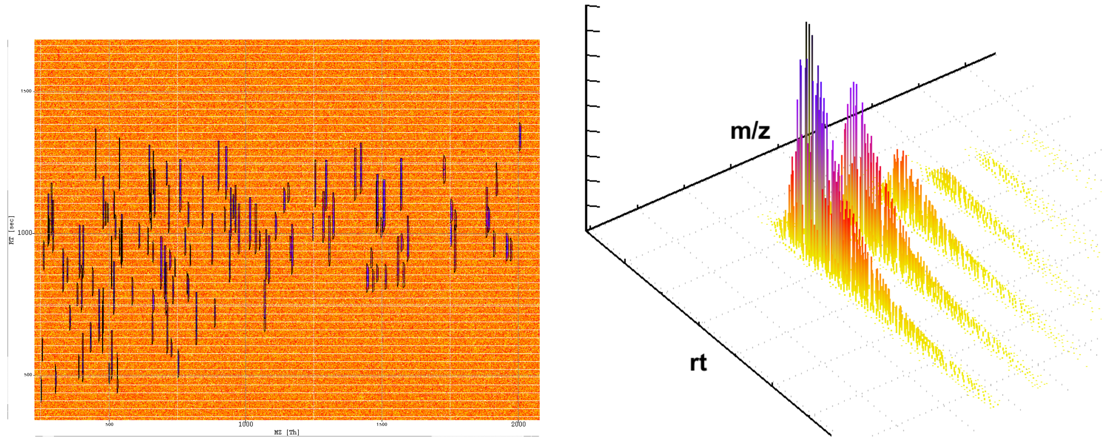
(b) Simulated elution peaks

Figure 4.4: *Comparison of real and simulated elution peaks: (Left) Real elution peaks from a reversed-phase HPLC experiment. (Right) Simulated elution profiles. They represent the three pre-defined column configurations* LC-MSSIM *can simulate.*

Putting all this together, we can model LC-MS experiments with different mass resolutions and chromatographic conditions. To exemplify this,Fig. 4.5(b) gives a bird's eye view of an LC-MS map created by LC-MSSIM. This map represents a tryptic digest of BSA (Bovine Serum Albumin) with some contaminations (metabolites etc., see below). Figure 4.5(a) shows a simulated BSA peptide ion from this map. This peptide occurs in the charge variant +1 and +2.

Addition of Noise and Contaminations

No real LC-MS data set consists only of true signals i.e. signals caused by sample compounds. There is always some (and often a high amount of) noise in each spectrum. LC-MSSIM has several parameters that allow the user to introduce noise of various forms into a data set. Users



(a) LC-MS map of the BSA digest

(b) A simulated BSA feature

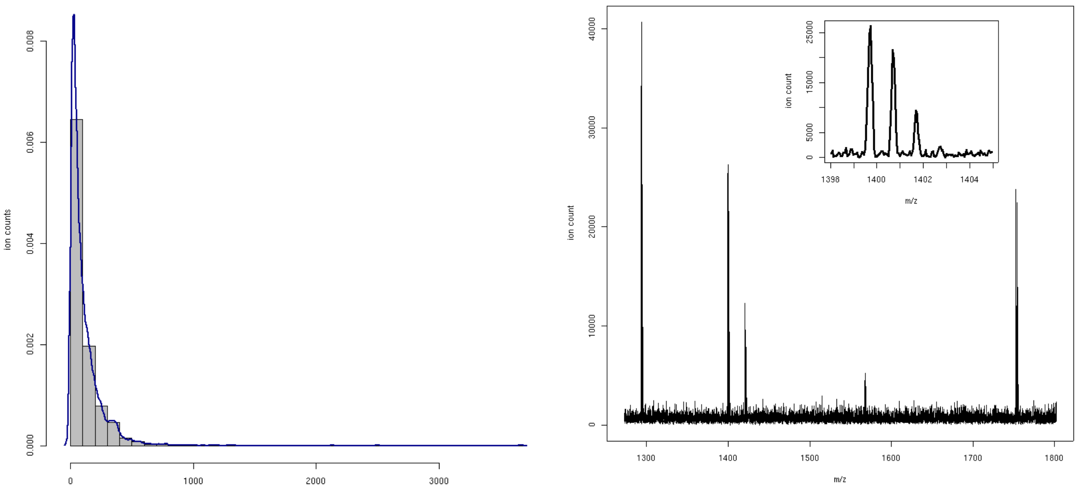
Figure 4.5: *Left: a bird's eye view of an LC-MS map of a simulated BSA digest. The blue signals are the tryptic peptides, red and yellow is shot noise and contaminants. Right: a simulated BSA peptide feature.*

can simulate almost perfect LC-MS runs and runs with high amount of noise posing severe challenges to data analysis algorithms.

First, the user can define error bounds on the theoretically predicted retention times. By doing so, we simulate retention time shifts between different experiments and, for instance, can evaluate the performance of LC-MS alignment algorithms that are used to correct for these shifts. LC-MSSIM assumes these errors to be Gaussian-distributed and the user can define medium and standard deviation in each case.

Mass analyzers with different mass accuracies and resolutions are simulated by changing the FWHM of the peptide peaks as described above and by altering the sampling step size of the peptide models. Furthermore, LC-MSSIM simulates inaccuracies in peak intensity measurements by adding Gaussian-distributed noise to peptide peaks. Finally, ESI mass spectra frequently contain high-frequency noise signals of low to medium intensity, often referred to as *shot noise*. This term stems from electronics and physics (R. Sarpeshkar and Mead, 1993) and describes statistical fluctuations occurring if the number of particles measured by a detector is very small. Its strength increases with the average intensity of the detected signal but is usually only detectable if the measured signal is very weak. The common assumption is that shot noise is Poisson-distributed (van Etten, 2006).

To our knowledge, the notion of shot noise in mass spectrometry is much less well defined than in physics but usually loosely refers to high-frequency noise of low intensity in a mass spectrum. Noise models for mass spectra have been the topic of several publications, but no consensus on the most suitable model exists so far (Anderle *et al.*, 2004; Du *et al.*, 2008; Shin *et al.*, 2004). However, recent publications suggests that noise in both Q-TOF and Ion Trap spectra can be modeled using a Poisson distribution (Du *et al.*, 2008) and therefore we decided to do the same. We split each spectrum in our simulated LC-MS map into segments of uniform size. We determine the number of shot noise signals per segment by sampling from a Poisson distribution, though m/z and intensity of these particles are given by a Gaussian and Exponential distribution, respectively. Figure 4.6(a) shows the peak intensity distribution of a *real* MS spectrum. The distribution is approximately exponential with some signals (true peptide peaks) having a high intensity. This shows that our model with exponentially distributed noise intensities well approximates real signals.



(a) Intensity distribution of a real mass spectrum

(b) A simulated spectrum

Figure 4.6: *A single simulated mass spectrum and its intensity distribution. Left: the intensity distribution (histogram and density plot) of raw data point intensities in a real spectrum. Right: A simulated mass spectrum of a BSA digest with shot noise added. The inset shows an isotope pattern from the same spectrum.*

Another typical phenomenon in mass spectra is a so-called baseline signal which usually decays with increasing m/z. This is usually a problem for MALDI instruments but less in ESI mass spectrometry. LC-MSSIM can simulate baseline signals by adding an exponentially-decaying baseline to each mass spectrum, but this feature is turned off by default.

Shot noise and a baseline are both factors that hamper a computational analysis. But of equal concern for feature detection algorithms are non-peptidic contaminations in an LC-MS experiment or peptide signals arising from modified peptides. Hoopmann *et al.* (2007) demonstrated that the detection of modified peptides is difficult and requires additional computational effort since the isotope pattern of these peptides does not follow the typical averagine pattern assumed by most algorithms. In short, an averagine is an average amino acid with a composition estimated from a large number of protein sequences. Using the averagine, we can estimate the average isotope pattern for a peptide of a given mass. Furthermore, contaminations such as salt molecules or metabolites are of lesser interest in proteomics studies and should not be reported by peptide feature detection algorithms. For these reasons, we decided to simulate these interferences as well. LC-MSSIM comes with a list of sample contaminants that can easily be extended by editing the corresponding text file. The current list of available contaminants comprises a snapshot of metabolites downloaded from the Human Metabolome Database (Wishart *et al.*, 2007). The user can set the percentage of added contaminants with respect to the number of peptides.

LC-MSSIM also includes a list of typical modifications such as oxidations or demethylations together with a list of affected amino acid residues. For each peptide containing a matching amino acid, LC-MSSIM determines at random whether the amino acid is modified or not. The user can set the corresponding probabilities and desired relative frequencies of modified peptides. Figure 4.6(b) shows an MS spectrum from a simulated BSA digest with added metabolic contaminants and shot noise. It is only a single spectrum from an LC-MS experiment, therefore not all BSA peptides are visible.

Quality of Simulation

Performing simulations always raises the question whether the simulated data is sufficiently close to reality. In this section, we will demonstrate that our simulations are realistic.

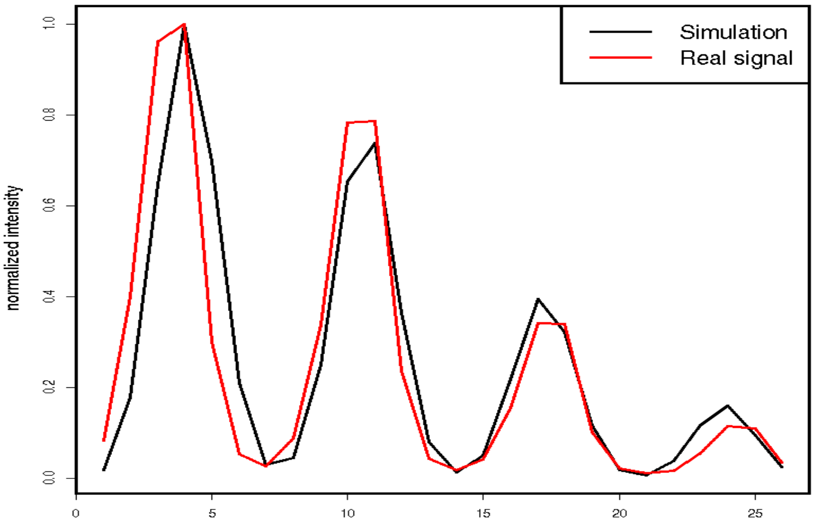
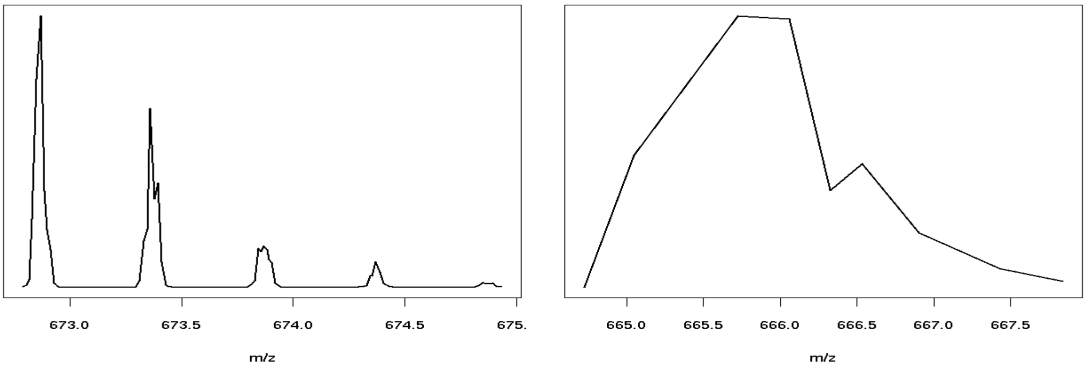


Figure 4.7: *Comparison of a simulated and a real isotope pattern. The Spearman correlation between these pattern is high:* 0.91

We already showed that our model of elution peaks and shot noise match real data well (see Fig. 4.4 and Fig. 4.6, respectively). To illustrate the quality of our isotope peak model, we simulated a mixture of standard proteins and generated a real LC-MS run of the same mixture on an ESI-TOF mass spectrometer (microTOF, Bruker Daltonics). Details of the sample preparation are given in Schley *et al.* (2006). We manually extracted peptide feature signals from the real data set and the simulated LC-MS run and computed Spearman's rank correlation coefficient for three simulated isotope peak patterns. The correlation coefficients were high, namely 0.91, 0.90 and 0.84. Figure 4.7 gives an example. We repeated this experiment with a low resolution LC-MS run. We recorded a mixture of human serum on an ESI ion trap instrument and simulated an LC-MS map of similar resolution. Details of the sample preparation are given in Mayr *et al.* (2006). The correlation between real and simulated isotope pattern was high (between 0.92 to 0.98). This shows that our isotope distribution model based on the algorithm by Kubinyi (1991) and a Gaussian peak shape generates realistic signals.

Influence of Mass Resolution

We downloaded the Mouse IPI protein sequence set (08.04.2008) and randomly selected 100 protein sequences from this set. A tryptic digest and filtering for detectability at a threshold of 0.8 resulted in 820 peptides. The chosen threshold corresponds to a False Discovery Rate of 10%. We opted for this mixture of moderate complexity to avoid a high number of overlapping peptides. Still, manual annotations of all these data sets would be tedious. In our first experiment, our goal was to determine to what extent the performance of current feature detection algorithms depends on the mass resolution of the instrument. We simulated different mass resolutions by changing the FWHM of the peptide isotope pattern. We generated data



(a) FWHM 0.05

(b) FWHM 0.8

Figure 4.8: *Left: a simulated isotope pattern with FWHM* 0.05*. Right: a simulated pattern with FWHM* 0.8*.*

sets for FWHM values of 0.05, 0.2, 0.5 and 0.8. A peak FWHM of 0.05 roughly corresponds to an Orbitrap instrument whereas an FWHM of 0.8 results in spectra similar to typical ion trap measurements. Figure 4.8 shows two exemplary isotope pattern generated for FWHM 0.05 and 0.8.

To each data set, we added shot noise with a mean intensity of 150 and a Poisson rate of 450. This noise level was chosen such that all peptide signals would be well above the noise level. The challenge of this benchmark was to detect poorly resolved and possibly overlapping peptide signals. The results of each algorithm differ depending on the parameters chosen. We describe our strategy to find suitable parameters for each algorithm in the appendix.

We noticed early on that each algorithm follows a different strategy. Some algorithms report a lot of potential peptide features even for simple data sets, rather than missing an important signal. The rationale seems to be that it is better to obtains many false positives than to miss a potentially crucial signal. Of course, spurious noise signals can be removed during later stages of the workflow. For instance, by removing signals that do not appear at consistent positions during alignment. Nevertheless, this makes matters unnecessary difficult. In contrast, some algorithms are highly specific but tend to miss poorly resolved signals. Which strategy is best might depend on the specific task to be performed and the complexity of the data.

Furthermore, not every algorithm associates a quality or confidence measure with a feature that could be used as a cutoff. It is therefore not possible to give the classical Receiver Operating Characteristic curves frequently used when comparing signal detection methods. Consequently, we decided to give the results in terms of the false discovery rate (FDR) and true positive rate (TPR). We approximate the FDR by

$$\text{FDR} = \frac{\text{False Positives}}{\text{True Positives} + \text{False Positives}}$$

and compute the TPR as

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$

We count a peptide signal as detected correctly if the algorithm found a feature with the correct monoisotopic m/z (with a tolerance of 0.8 m/z) and an estimated bounding box within the true signal bounding box. It happens frequently that an algorithm splits a feature eluting over a longer period of time into several parts, i.e., loses track of the elution peak. In this case, we counted only one true positive hit for this feature but did not count the remaining features as incorrect hits.

FWHM	Decon2LS	msInspect	MZmine	SpecArray	Superhirn	SweepWavelet
0.05	0.80/0.64	0.96/0.46	0.78/0.87	**0.99**/**0.05**	0.88/0.57	0.92/0.27
0.2	0.82/0.78	0.32/0.97	0.71/0.83	**0.97**/**0.41**	0.68/0.50	0.96/0.05
0.5	0.68/0.49	0.80/0.98	0.33/0.68	0.94/0.54	0.10/0.72	**0.95**/**0.02**
0.8	0.49/0.34	0.87/0.98	0.03/0.31	**0.92**/0.36	0.003/0.50	0.84/**0.12**

Table 4.4: *Comparison of all feature detection algorithms on simulated LC-MS maps with different mass resolutions: The first entry in each column gives the true positive rate, the second one is the false discovery rate.*

Table 4.4 shows the results of this experiment. All algorithms recover most of the peptide signals at high mass resolutions but the true positive rate decreases for all algorithms with the resolution. SpecArray performs best on the high resolution data but with decreasing performance at lower resolutions. We would like to emphasize that SpecArray performs very well out-of-the-box, i.e. without parameter tuning, whereas other algorithms required a filtering of signals. Our algorithm, SweepWavelet, is much less affected by a poor mass resolution than all other algorithms but fails to detect some signals even at high resolutions. SweepWavelet also gives the on average lowest false discovery rate on all data sets.

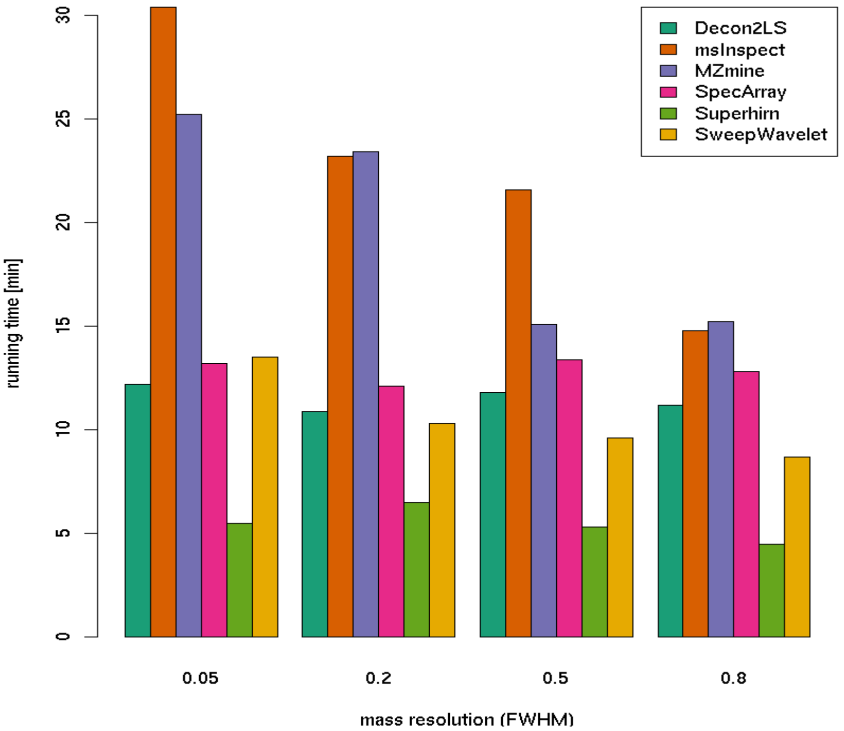


Figure 4.9: *Running times of all feature detection algorithms on the four data sets with different mass resolutions. Although all running times are within an acceptable range, the software* SpecArray *is the fastest.*

We also note that some algorithms, especially msInspect and Decon2LS, compute huge numbers of false positives and consequently, their False Discovery Rates are poor. We made the same observation already in the previous section. On the other hand, both algorithms find almost all true signals, especially on the high resolution data set.

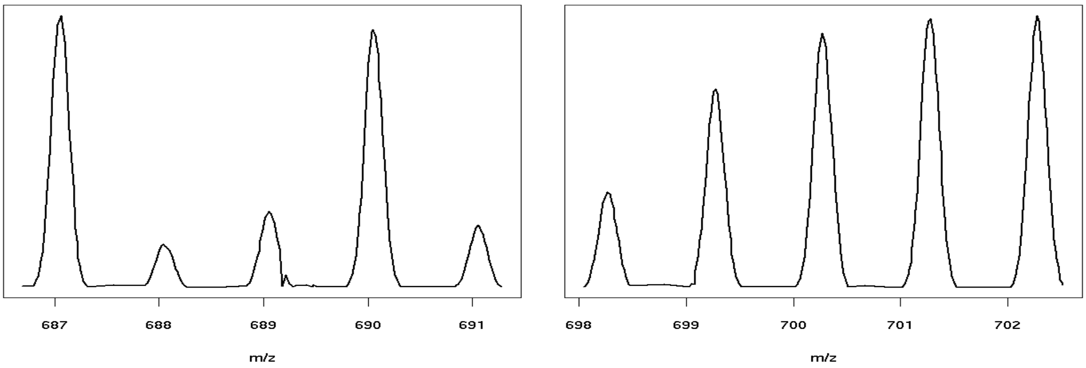
Figure 4.9 displays the running times of all algorithms on each data set. The time measurements were performed on a 3.2 GHz Intel Xeon CPU with 3 GB memory running Debian or Windows Server 2003 R2 (in the case of Decon2LS, MZmine and msInspect). Note that the running times of Decon2LS and MZmine are approximate results only, since both tools are GUI-based and therefore do not allow direct time measurements. SpecArray stands out as the fastest algorithm, whereas all other tools yield acceptable running times. In our point of view, feature detection should not take longer than the acquisition of a typical LC-MS map, and this is achieved by all algorithms.

To summarize, different algorithms have different strengths: some recover nearly all true signals even under poor conditions but at the expense of large numbers of false positive hits. One might argue that many of this false positive signals could be removed by removing features of low intensity or of unlikely masses. But this clearly has its disadvantages if we examine complex mixtures with large dynamic ranges and many compounds at low intensities.

Decoy Isotope Pattern

Another important characteristic of a feature detection algorithm is its ability to discriminate between peptide isotope pattern and signals caused by other compounds such a column material or any other non-peptidic substance in the sample. Declaring these signals incorrectly as peptide feature will hamper the further downstream data analysis.

To test the discriminatory power of each feature detection algorithm, we simulated two LC-MS maps with decoy isotope pattern. For the first map, we replaced the intensity of each isotope peak by an intensity sampled from a uniform distribution (*uniform decoy*). For the second map, we replaced the intensity i of each peak by $1 - i$ (*inverse decoy*). All peaks were simulated with FWHM 0.2. Both simulations should yield signals being very different from true isotope pattern. Whereas the uniform decoys might exhibit some similarity to true isotope pattern by chance, the inverse decoys should not.



(a) Uniform decoy

(b) Inverse decoy

Figure 4.10: *Left: isotope peak heights sampled from a uniform distribution (uniform decoy). Right: isotope peak i replaced by $1 - i$ (inverse decoy). All peaks were simulated with FWHM* 0.2.

Figure 4.10 shows an example of a uniform and an inverse decoy isotope pattern. The uniform decoy map contains 361 decoy isotope pattern and the inverse decoy map 359 pattern. For each algorithm, we counted how many pattern this algorithm would declare as true isotope pattern. Table 4.5 contains the results. Only Superhirn and SweepWavelet yield acceptable results and declare only a small percentage of decoys as true features. SweepWavelet detects by far

data set	Decon2LS	msInspect	MZmine	SpecArray	Superhirn	SweepWavelet
inverse	55.71% (72294)	10.51% (615)	100.00% (137512)	10.86% (42)	0.58% (64)	**0**% (0)
uniform	54.01% (64038)	20.61% (1005)	99.72% (69115)	88.92% (646)	0.14 (152)	**0.01**% (24)

Table 4.5: *The percentage of uniform and inverse decoy pattern that were declared as true isotope pattern by each algorithm. The number in brackets give the overall number of features. The uniform decoy map contained* 361 *decoy pattern and the inverse decoy map* 359 *pattern.*

the smallest number of features in the decoy data set. All other algorithms do not discriminate well. MZmine stands out as declaring almost all decoys as true features.

Metabolite Decoys

Metabolites are likely to be the second most abundant class of sample compounds after peptides. In this experiment, we tested to what extent current peptide feature detection algorithms can discriminate between peptide signals and signals of other sample compounds. To this end, we generated an LC-MS map consisting of 360 metabolites, but no peptides. These metabolites represent a random subset of compounds from the Human Metabolome Database (accessed 11 March 2008). For each metabolite, we computed its isotope distribution and placed it at a randomly-determined retention time in the LC-MS map. We modeled the elution profile using a Gaussian function.

	Decon2LS	msInspect	MZmine	SpecArray	Superhirn	SweepWavelet
PF (%)	99.17	99.17	**80.44**	98.90	98.07	92.56
# features	3722	5885	283	719	655	813

Table 4.6: *Percentage of metabolites declared as features. We also report the number of features detected in total.*

Table 4.6 shows the results of this experiment. The row labelled with PF indicates the percentage of metabolite compounds declared as peptide feature by the algorithm. The second row gives the total number of features reported. As we can see, most algorithms are not able to distinguish between peptide and metabolite isotope pattern. This is probably not surprising as peptides and metabolites exhibit similar isotopic patterns. To illustrate this fact, we drew 2875 metabolites at random from the Human Metabolome Database and computed their isotope pattern. After this, we estimated the average isotope pattern for a peptide of the same mass using the method of averagines Senko *et al.* (1995a) which was explained in the previous chapter. This method is used by most feature detection algorithms developed so far.

We computed the Spearman rank-based correlation coefficient for both, metabolite and peptide isotope profiles. The lowest correlation is 0.96, which is still high. But this means that current algorithms that try to detect peptide signals using the averagine method will only poorly be able to distinguish peptides from other biomolecules.

This problem might not be grave. If we simply search for signals that discriminate between two conditions e.g. control and disease, it might at first not be that important whether this signal is caused by a peptide or a metabolite. But it is a fact that users have to keep in mind: most feature detection algorithms detect a lot of features in a real world data set, many more than are sequenced. This has usually been attributed to the fact that the Data-dependent Acquisition (DDA) process is a semi-random sampling of sample compounds and many peptides will never be identified. But users need to be aware that not all detected features will be caused by peptides, but also by other biomaterials including metabolites.

Influence of Chromatographic Condition

condition	Decon2LS	msInspect	MZmine	SpecArray	Superhirn	SweepWavelet
good	0.72/0.38	0.97/0.85	0.85/0.19	**0.98**/**0.05**	0.85/0.20	**0.98**/0.87
medium	0.81/0.64	0.98/0.84	0.92/0.28	**0.99**/**0.05**	0.92/0.23	**0.99**/0.87
poor	0.79/0.48	0.98/0.84	0.91/0.20	**0.99**/**0.05**	0.92/0.16	0.98/0.86

Table 4.7: *True positive (first entry in each column) and false discovery rates (second entry) on three data sets under changing chromatography conditions.*

We simulated three LC-MS runs, one for each predefined chromatographic condition (good, medium and poor) but kept the mass resolution in terms of the FWHM constant at 0.05. If a peak elution profile gets noisy, we expected most algorithms to lose track of the isotope pattern over time or maybe even not to detect it at all.

Table 4.7 shows the results of this experiment, again in terms of False Discovery and True Positive Rate. The performance of most algorithms remains stable across chromatographic conditions. There are only two algorithms whose performance lags behind if the elution peaks become noisier, SweepWavelet and MZmine. The first simulated run with good column conditions contains many overlapping isotope pattern, and SweepWavelet is not able to separate strongly overlapping signals. Furthermore, SweepWavelet uses a Gaussian model to fit the elution curve of a feature and discards features having a poor probability under this model. Obviously, this dampens SweepWavelet's performance in this experiment. MZmine does not seem to perform well on high resolution data in general, as we already observed in the previous section. The false discovery rate of SpecArray increases slightly at poorer chromatography conditions. All other algorithms are not affected. Especially Decon2LS is not affected by changes in the chromatographic condition since this tool detects isotope pattern in a spectrum-wise matter and does not take the elution profile into consideration.

4.3 Summary and Discussion of Results

In the previous chapter, we presented a novel algorithm, SweepWavelet, for the quantification of peptide signals in LC-MS maps. In this chapter, we evaluated its performance on real and simulated data sets and compared it to other recently published algorithms.

All algorithms show good performance, but with significant differences depending on data set and task. MZmine performs well on low resolution spectra such as the ones from the myoglobin study. It has weaknesses if peptide ions of higher charge states appear in the data since it does not consider charge states larger than three. It also tends to pick up too many noise signals. This is also the disadvantage of Decon2LS which re-implements the THRASH algorithm and produces very high false discovery rates.

Superhirn and SpecArray are tuned for high-resolution data. They give excellent results on LC-MS maps recorded on modern mass spectrometer such as the peptide mixture or the simulation performed for low FWHM values. Superhirn is likely to produce too many false positives. High-resolution mass spectrometer will at some point be widely available but right now, low resolution instruments are still the work horses in most laboratories. It is therefore important to have methods available that work well on low resolution data, too. The quantification performance of both algorithms is less satisfying on lower resolution data.

MsInspect and SweepWavelet both use wavelets to detect potentially interesting peaks before combining them into peptide features. However, our approach SweepWavelet has some advantages: it performs well also on low-resolution data where MsInspect has its weaknesses as we demonstrated on the myoglobin data set. The isotope wavelet and our peptide template model account for poor mass resolutions and are well able to recover weak isotope patterns. Furthermore, our algorithm makes more assumptions about interesting signals by incorporating the relative peak intensities directly into the mother wavelet. This leads to fewer false positive features as in the case of the peptide mixture LC-MS data and of the simulated decoy maps. It seems that our algorithm is not the best on every data set but widely applicable and, importantly, only marginally influenced by mass resolution and quality of the data. It also yields very few false positives and is well able to discriminate between true and decoy isotope pattern.

Nevertheless, our algorithm has also room for improvement. Since we compute one instance of the isotope wavelet for each considered charge state, the user has to make some assumptions about the charge states that will appear in the data. Computing the isotope wavelet for all charge states in a large range, say from 1 to 20 for whole protein mass spectrometry, will increase its running time. Our simulations also revealed that SweepWavelet does not perform well if the chromatographic conditions are poor, i.e. if the elution peaks do not fit its assumptions well. In future revisions, one might consider to remove the elution peak model and only filter for the m/z part of the model. On the other hand, one might argue that features with very noisy elution curves might not be reliable measurements anyway and there is no harm in discarding them.

We made an effort to test all algorithms in diverse scenarios and on different data sets. Nevertheless, there is still a lack of benchmark data sets and suitable performance metrics for the comparison of peptide detection and quantification algorithms. The LC-MS data deposited in public databases so far is either of poor quality or does not allow for a meaningful comparison of quantification algorithms since the contained peptides and their quantities are not known (e.g. for whole cell digests). Ideal test cases would be manually annotated data sets of high complexity and quality but these are of course expensive and difficult to obtain.

Chapter 5

Statistical Quality Assessment of LC-MS Experiments

Synopsis: *This chapter presents methods for the statistical quality assessment of large scale LC-MS experiments. Controlling data quality in these experiments is crucial since reliability and accuracy of the results obtained are strongly influenced by it. Nevertheless, methods for quality control that are widespread in other scientific fields, are not very common in mass spectrometry-based proteomics. We present novel methods to detect outliers among large sets of LC-MS runs. Our method is based on quality descriptors capturing different aspects of the quality of an LC-MS run and robust statistics that help us to define accurate confidence levels for outlier detection.*

5.1 Introduction

This chapter addresses the problem of quality assessment in large scale quantitative LC-MS studies. So far, this is a relatively unexplored topic. There are, however, some publications on the quality assessment of MS fragmentation spectra (Bern *et al.*, 2004; Choo and Tham, 2007; Na and Paek, 2006; Moore *et al.*, 2000; Xu *et al.*, 2005; Flikka *et al.*, 2006). But their focus is different: the aim of these methods is to detect and remove low quality MS/MS spectra from an LC-MS/MS run. The rationale is that these spectra would not be annotated by identification algorithms anyway and that their removal will lead to a significant speed-up of the data analysis. As we described in Chap. 2, fragmentation and sequencing of the peptides using MS/MS is an additional experimental step and is not the focus of this work. A shorter version of this chapter is submitted for publication (Schulz-Trieglaff *et al.*, 2008c).

The analysis of LC-MS maps is a sophisticated task and requires several computational steps such as denoising, peptide feature detection, alignment and statistical analysis as we explained in Chap. 2. After differentially expressed peptide features have been found, they need to be sequenced using MS/MS-based identification and have their abundances and sequences mapped to the parent protein. These are general steps which usually have to be adopted depending on the aim of the study. But each of these computational steps has its own difficulties and a typical workflow is complex and error prone (Stead *et al.*, 2008). It is therefore desirable to identify poor LC-MS runs as early as possible. This would us allow to either exclude outlier runs from the further analysis, to repeat them or at least to downweight these measurements to reflect our reduced confidence.

The method that we present controls the data quality in a quantitative LC-MS experiment. Numerous problems can affect the quality of an LC-MS run. Among these are: instabilities of the chromatography, uncontrolled degradation of the peptides or artifacts in the mass spectra caused by the LC mobile phase or buffer molecules. All these problems will have significant influence on the quantitative data that is generated in later steps of the workflow and hamper the downstream computational data analysis.

Only little work on the quality and reproducibility assessment of mass spectrometry data has been published so far (Coombes *et al.*, 2003; Prakash *et al.*, 2007; Whistler *et al.*, 2007). Prakash *et al.* (2007) use a distance measure computed by their alignment algorithm to highlight problems of reproducibility in several LC-MS data sets. Their method is successful in visualizing the time order in which the LC-MS runs were performed and reveals patterns caused by changes of the chromatographic column or instrument settings during a study. But their method does not provide direct information on outlier runs and when to discard them but visualizes only general trends in the data. Both Whistler *et al.* (2007) and Coombes *et al.* (2003) address the problem of noise filtering, quality assessment and outlier removal but focus solely on SELDI-TOF spectra which are one-dimensional and thus much less complex than LC-MS maps.

In contrast to mass spectrometry-based proteomics, quality assessment and control methods are more established in gene expression studies (Brettschneider *et al.*, 2007; Brown *et al.*, 2001; Cohen Freue *et al.*, 2007; Model *et al.*, 2002). Brettschneider *et al.* (2007) review the state of the art of statistical quality control and present two methods for the quality assessment of Affymetrix arrays. They exploit the fact that Affymetrix arrays contain several oligo probes for each gene. The standard approach is to combine the oligo signal intensities for a gene using a regression model (Irizarry *et al.*, 2003). This model is learned from the intensity values for a sample with known composition and abundances. They devise different summary statistics based on the residuals of their regression model and show that outlier arrays can be identified by comparing and visualizing these quality values across different arrays. Brown *et al.* (2001) use image features of single slides to find poor quality microarrays in a batch of experiments. They perform a pixel-by-pixel analysis of individual spots on the array to estimate background noise and fabrication artifacts. They derive a weighting scheme based on these estimates and demonstrate that they can improve the accuracy of the expression measurements using it. Cohen Freue *et al.* (2007) apply the Mahalanobis distance to detect outlier runs in large-scale gene expression studies using Affymetrix arrays. They represent each array by a set of quality values. These values are automatically computed by the image acquisition software and among them are the percentage of spots present (e.g. active) on the array, the noise level or background intensity. They show that they can accurately detect outlier arrays that were previously confirmed by other quality control methods. Finally, Model *et al.* (2002) borrowed methods from the field of statistical process control. They treat a set of microarray measurements as a statistical process, define a subset of arrays (usually from the beginning of the study) as reference or *historical data set* and compare all subsequent arrays to this reference. If the distance of an array to the historical set becomes too large they declare this array as being out of control. Model *et al.* (2002) demonstrate that they can detect accurately changes in temperature or a saturation of signal intensity. Of course, this method requires the identification of a good and high quality reference subset as all following arrays are compared to this reference set. Obviously, it is not always easy to identify such a subset but the accuracy of this method depends strongly on it.

In this work, we investigate how established methods from the field of statistical quality control can be extended and applied to LC-MS data. We developed a set of methods for the identification of outlier LC-MS maps from a set of replicates. Our approach is based on sound statistical principles and we demonstrate that we can accurately detect dubious LC-MS runs in two large scale studies.

5.2 Mathematical Preliminaries

Before we present our method, we need to introduce some mathematical terms that we will use during the later course of this chapter. We will represent an LC-MS map using n numerical quality descriptors e.g. each map will be represented by a vector x[1] where each entry contains the value of a descriptor.

5.2.1 Fundamental Statistics

From a statistical point of view, these vectors are realizations of a *multivariate random variable*, obtained from an underlying distribution by a sampling process. These realizations are called samples or random variates. We can estimate several statistical properties of the underlying distribution from a sample. In mathematics, an *estimator* is a function of the sample that we use to estimate an unknown distribution parameter. The *bias* of an estimator is the difference between its expected value and the parameter being estimated. An estimator is unbiased if this difference equals zero.

To give an example, the *sample mean* $\hat{\mu}$ of a set of observations $\{x_1, x_2, \ldots, x_n\}$ is given by

$$\hat{\mu}_X = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5.1}$$

The sample mean is an unbiased estimator for the population mean μ_X. Note that for a multivariate sample, the sample mean is a vector where each entry is the mean of the corresponding

[1] Remark: in the following, we will use lower case letters for (column) vectors and capital letters for matrices.

sample coordinates. Therefore we might also speak of a coordinate-wise mean if we want to emphasize the geometric point of view.

The *variance* of a sample is defined as

$$\hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n}(x_i - \mu_X)^2}{(n-1)} \tag{5.2}$$

The variance is a measure of spread or scatter of the sample. It can be interpreted as the sum of squared distances of each point to the mean μ_X. We divide by $n-1$ instead of n to obtain an unbiased estimator.

The *covariance* of two samples, each with sample size n, is defined as

$$cov_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{(n-1)} \tag{5.3}$$

The covariance gives a measure of the amount of correlation between the two sets of random variates. It is zero for uncorrelated variates, positive if both X and Y increase together and negative if X increases when Y decreases. Note that if $X = Y$, the covariance is equal to the variance.

Finally, the covariance matrix is the generalization of variance and covariance for multivariate samples. It stores the variances and covariances for all dimensions of the sample. The entries on the diagonal of the matrix are the variances, entries off the diagonal are the covariances:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & cov_{12} & \dots & cov_{1m} \\ cov_{21} & \sigma_{22}^2 & \dots & cov_{2m} \\ \vdots & & & \\ cov_{m1} & cov_{m2} & \dots & \sigma_{mm}^2 \end{pmatrix} \tag{5.4}$$

where σ_{11}^2 is the variance of the first dimension, cov_{12} is the covariance between the first and second dimension and so on. Since $cov_{X,Y} = cov_{Y,X}$, the covariance matrix is symmetric.

Σ is also positive semi-definite. A matrix M with dimension $n \times n$ is positive semi-definite if $x^T M x \geq 0$ for all vectors $x \in \mathbb{R}^n$ and $x \neq 0$. Σ can be mean-centered and written as $\mathbb{E}(X^T X)$. To see that Σ is positive semi-definite, consider that

$$v^T X^T X v = (vX)^T (Xv) \geq 0 \tag{5.5}$$

for any real-valued vector v. Since the expectation value of a non-negative random variable is also non-negative, it holds that

$$v^T \mathbb{E}(X^T X) v = \mathbb{E}(v^T X^T, Xv) \geq 0 \tag{5.6}$$

and thus Σ is positive semi-definite.

To sum up, we introduce a concept from statistical hypothesis testing, called the *Bonferroni correction* for multiple testing (Bonferroni, 1935). In Chap. 3, we already used the correction by Benjamini and Hochberg (1995) to account for multiple testing problems in the scoring of peptide signals. Here, we also deal with a multiple testing problem since we will repeatedly perform statistical tests to check if an LC-MS map is an outlier or not. It might happen that a test results in a positive result by chance alone since we test so many hypotheses at the same time. But for our outlier detection problem, the multiple testing problem is less pronounced since we test only $60-100$ maps and not several 10.000 peptide features.

Consequently, we use the straightforward Bonferroni correction which simply adjusts the alpha value for a set of n tests to α/n. This is obviously a conservative adjustment which results in low values for alpha if n, the number of tests, is large. But for our purposes, this is not a problem. If our method is applied to larger sets of LC-MS maps, we would however recommend to use more sophisticated approaches such as the correction by Benjamini and Hochberg (1995).

5.2.2 Linear Algebra: Eigenvectors and Eigenvalues

In the later course of this chapter, we apply a method called *principal component analysis* (PCA). PCA relies on fundamental concepts from linear algebra, such as eigenvectors and eigenvalues. A non-zero vector x of dimension n is an *eigenvector* of a square matrix $A \in \mathbb{R}^{n \times n}$ if it satisfies

$$Ax = \lambda x \tag{5.7}$$

where λ is a scalar and called the *eigenvalue*. This equation is called the *eigenvalue equation*. In other terms, the linear transformation represented by A only changes the length of the vector x. This is equivalent to

$$(A - \lambda I)x = 0 \tag{5.8}$$

where I is the identity matrix. This corresponds to a set of homogeneous, linear equations. It has a non-trivial solution (e.g. $x \neq 0$) if and only if the determinant of A equals zero. This in turn yields an equation for the eigenvalues of A:

$$p_\lambda := \det(A - \lambda I) = 0 \tag{5.9}$$

p_λ is called the *characteristic polynomial* and the equation is the *characteristic equation*, a nth order polynomial in the unknown λ. The polynomial p_λ has different real-valued n_λ solutions where $1 \leq n_\lambda \leq n$. Each solution is an eigenvalue. We can find the corresponding eigenvectors by inserting the computed eigenvalues into Eq. 5.7. Note that the eigenvectors of different eigenvalues of a matrix are pairwise orthogonal. Furthermore, the eigenvalues of a positive semi-definite matrix (such as the covariance matrix) are always non-negative.

There exist straightforward algorithms to compute the eigenvalues and eigenvectors for small matrices. Since the matrices we deal with in this work are usually large, we use more sophisticated algorithms from the Linear Algebra PACKage (LAPACK) (Anderson *et al.*, 1999) which are included in the R software (http://www.r-project.org).

5.3 Methods

In contrast to computational biology, methods for quality control are widely spread in other fields, such as manufacturing. These methods usually define a number of variables for a product whose quality is to be monitored. As an example, a company that manufactures bolts would monitor size, diameter and milling of the bolt and would like to ensure that these variables remain within a given tolerance for each bolt produced. In this work, our product to monitor are raw LC-MS maps.

By “raw”, we mean the unprocessed spectra before any noise filtering, peak detection or centroiding has been performed. Most statistical methods for quality assessment expect that each item is described by one (univariate) or several (multivariate) variables. For LC-MS maps, it is not clear what suitable variables could be. One straightforward approach is to describe an LC-MS map by all its data points. But the number of data points (not peaks) in an unprocessed

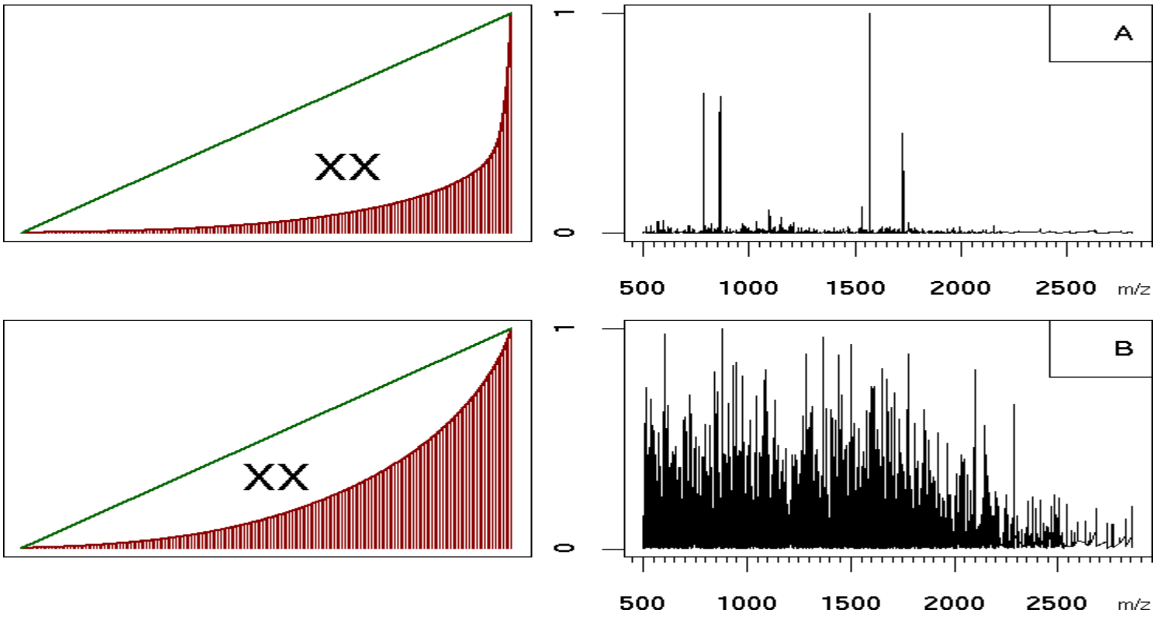


Figure 5.1: *Two examples of the Xrea value: spectrum A is of good quality with low background noise where spectrum B has a large amount of noise. The plots to the left show the corresponding Xrea distribution. The area denoted by XX is used in Equation 5.11.*

LC-MS map is huge, easily several millions of points. Second, many of the raw data points in a map will be caused by noise and might distort the results of an automatic outlier detection.

Consequently, we devised a list of *quality descriptors* to describe an LC-MS map. Some of these descriptors were taken from the literature, where they have been shown to be useful criteria for spectra mining and filtering tasks. Other descriptors were developed by us. Using these quality descriptors, we can now describe each map as a vector x and apply statistical methods to detect runs of poor quality.

We emphasize that we define quality in terms of reproducibility, i.e., an LC-MS map is of poor quality if its quality descriptors differ significantly from the descriptors of the other maps. It is thus important to compare only maps that represent the same subsets of a sample. As an example, in a multidimensional chromatography experiment, we can only compare LC-MS recordings of the same chromatography steps.

The programs to compute the quality descriptors for an LC-MS map were written using OpenMS (Sturm *et al.*, 2008), our software library for computational mass spectrometry. We performed the statistical analysis and visualization of the results using the mathematical software package R (www.r-project.org).

5.3.1 Quality Descriptors

We use a set of quality descriptors to an LC-MS map. These descriptors capture various aspects of the map, such as peaks and noise level of the spectra, as well as shape and reproducibility of the TIC. The descriptors are:

- Median of the Euclidean distances $D_E(s, s') = \sqrt{\sum(s_i - s_i')^2}$ between baseline-removed spectrum s' and original spectrum s for all spectra. The baseline or background noise in a mass spectrum is usually caused by molecules from the mobile phase of the column. Spectra with a large amount of background noise are difficult to analyze automatically. This rationale of this descriptor is that spectra with a strong baseline signal will be very different after baseline removal and thus have a large distance D_E. We remove the baseline using a TopHat-Filter which is a standard method for this task. To obtain a meaningful distance measure between the spectra, we perform a resampling of the mass spectrum and thus deal with a uniform spacing.

- Median of the Euclidean distances D_E between smoothed spectrum and original spectrum for all spectra. Consequently, a noisy spectrum will exhibit a large distance D_E to its smoothed version. We performed the smoothing using a Gaussian Filter with a kernel width of ≈ 2.0, depending on the mass spectral peak width. These two quality descriptors were firstly suggested by Windig *et al.* (1996) but they applied them to chromatograms to remove noisy mass traces from the LC-MS map.

- The *Xrea* value. This measure for the quality of a mass spectrum was already proposed by Flikka *et al.* (Flikka *et al.*, 2006). They developed it to filter MS/MS spectra before submitting them to a sequence database search. We will show that the *Xrea* criterion can equally be applied to MS spectra. The *Xrea* value is based on a cumulative intensity normalization. First, we normalize the spectral intensities by dividing by the total intensity. The *cumulative normalized intensity* of each data point in the spectrum is defined as the sum of the normalized intensities of all points with intensities smaller than or equal to the intensity of this point. Accordingly, the cumulative normalized intensity of the nth highest data point x is given by:

$$\frac{\sum(I(x)|\mathrm{Rank}(x) \geq n)}{\sum I(x)} \tag{5.10}$$

where $I(x)$ is the intensity of point x and Rank(x) represents the order of points if sorted by intensity in descending order. That is, the most intense point has rank 1, the next rank 2 etc. The numerator is divided by the sum of all intensities. This normalization is stable and not very dependent on the most intense peak which is a disadvantage of a normalization by the intensity sum. But in contrast to other methods such as a rank-based normalization, it does not discard the entire information in the spectral intensities.

We use the distribution of cumulative intensities to derive a measure for the quality of a spectrum. Figure 5.1 shows a plot of two spectra and their cumulative intensities. Spectrum A would usually be considered a good quality spectrum: several peaks with high intensity and only a low amount of background noise. Spectrum B is of poor quality: it contains a lot of noise signals and a strong background signal. To the left of each spectrum, we give the distribution of Xrea values. Its upper bound is the plot of each point versus its rank, indicated by the green line. We can see that, for spectra with a more uniform distribution, the distribution of cumulative intensities approximates the diagonal. For spectra with less background noise, but some pronounced peaks, the distribution is more peak-shaped. Note that the spectra to the right are MS and not MS/MS spectra.

Flikka *et al.* (2006) proposed to use the area between the diagonal and the cumulative intensity distribution as an indicator of the quality of a spectrum. In the plots in Fig. 5.1, this area is indicated by *XX*. If this area is large, the spectrum has low background noise and some elevated peaks of high intensity, which is desirable for feature detection and quantitation. If this area is small, all intensities in the spectrum are similar and the information content of this spectrum is rather low. To sum up, the Xrea quality descriptor is given by:

$$\mathrm{Xrea} = \frac{\text{area XX}}{\text{area of lower right triangle} + \alpha} \tag{5.11}$$

where α is a correction term to account for cases in which the highest point is significantly larger than the rest. We compute the area under the assumption that adjacent leg has unit length. Following Flikka *et al.* (2006), we set α to the relative intensity of the most abundant data point.

- Median of the number of data points with intensity ≥ 0 in each spectrum. This descriptor accounts for variations in the number of recorded intensities.

- Summary statistics for m/z, intensity and signal-to-noise ratio of all spectra. The summary statistics consist of minimum, maximum, mean and median. We estimate the noise level using an iterative sliding window approach. We move a window of size 25 Th across each spectrum and calculate a noise level for each window. We compute mean and standard deviation σ of all intensities in the current window and discard all points with an intensity higher than $3 \times \sigma$. We repeat this procedure and estimate the local noise level as the medium intensity after three iterations.

- Skewness and kurtosis of the TIC. For good and reproducible LC-MS runs, the TICs should exhibit similar shapes. Skewness and kurtosis describe the asymmetry and peakedness of a distribution, respectively. The skewness is the third standardized moment of a distribution. For a sample of size n, it is defined as

  $$\text{skew} = (1/n) \sum_{i=1}^{n} ((x_i - \bar{\mu})/\bar{\sigma})^3. \tag{5.12}$$

  The skewness is positive for distribution with a tail to the left, and negative for right-tailed distributions as illustrated in Fig. 5.2. The kurtosis is defined as

  $$\text{kurtosis} = ((1/n) \sum_{i=1}^{n} ((x_i - \bar{\mu})/\bar{\sigma})^4) - 3. \tag{5.13}$$

  It measures how sharply peaked a distribution is, relative to its width. We subtract 3 to achieve a kurtosis of zero for the Gaussian distribution. A distribution with positive kurtosis has more probability mass around the mean than the Gaussian distribution whereas a distribution with negative kurtosis has less probability mass around the mean and is therefore less peak-shaped. We give an example in Fig. 5.3.

- Minimum and maximum intensity of the TIC. We store the maximum and minimum intensity over the whole LC-MS run.

Using the descriptors described above, we can now represent an LC-MS map as a vector x where each entry of this vector represents one of the quality descriptor described above.

5.3.2 The Mahalanobis Distance

To decide whether an LC-MS map is an outlier compared to the rest of the measurements, we use the Mahalanobis distance (Mahalanobis, 1936). It has previously been applied to assess the quality of microarray experiments (Cohen Freue *et al.*, 2007) or for face recognition (Fraser *et al.*, 2003). The Mahalanobis distance is closely related to the Euclidean distance of two vectors x and y which is defined as:

$$D_E(x, y) = \sqrt{(x - y)^T (x - y)} \tag{5.14}$$

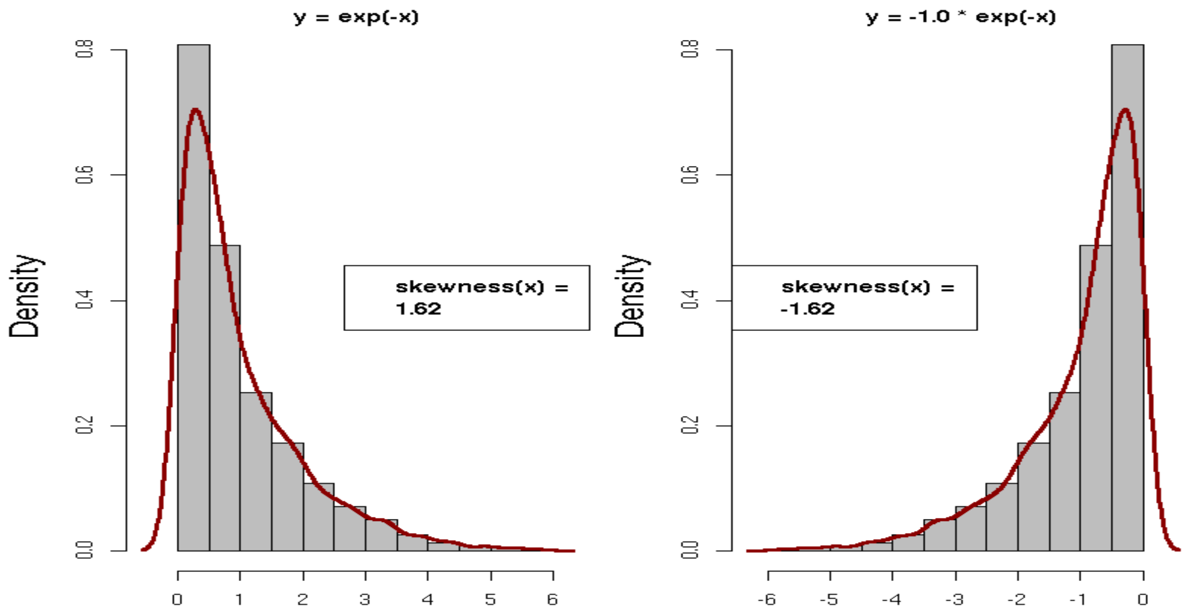


Figure 5.2: *The skewness of a distribution is positive if the distribution is right-tailed. It is negative if the distribution has a tail to the left.*

The square root is sometimes omitted for classification tasks since it does not affect the order of distances but requires an additional computational step. The Mahalanobis distance of two vectors x and y is defined as:

$$D_M(x,y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}. \tag{5.15}$$

Σ^{-1} is the inverse covariance matrix, also called the *precision matrix*. It differs from the Euclidean distance in the fact that it takes correlations of the data into account and weights each dimension by its variation. The Mahalanobis distance is equal to the Euclidean distance if the covariance matrix is the identity matrix e.g. if the individual dimensions are uncorrelated and have unit variance. To understand the difference between the Mahalanobis and the Euclidean distance, we need to recall that under the Euclidean distance, the set of points having the same distance to a given location is a sphere. The Mahalanobis distance stretches this ellipsoid to account for correlations and scales of the different dimensions.

Finally, we can also express the Mahalanobis distance in terms of a distance of a vector x to a distribution with mean $\mu = (\mu_1, \mu_2, \dots \mu_p)^T$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}. \tag{5.16}$$

This is the formulation that we will use in the later course of this chapter. Here, the mean μ and the covariance matrix Σ are determined by all LC-MS maps we consider. We use the Mahalanobis distance to classify maps having a large distance to the overall distribution as outlier.

Outlier Detection using the Mahalanobis Distance

Using the Mahalanobis distance, we can measure the distance of each LC-MS run, described by the vector of its quality descriptors x, to the distribution of all other n runs, characterized by their mean vector μ and covariance matrix Σ. The Mahalanobis distance of a vector with

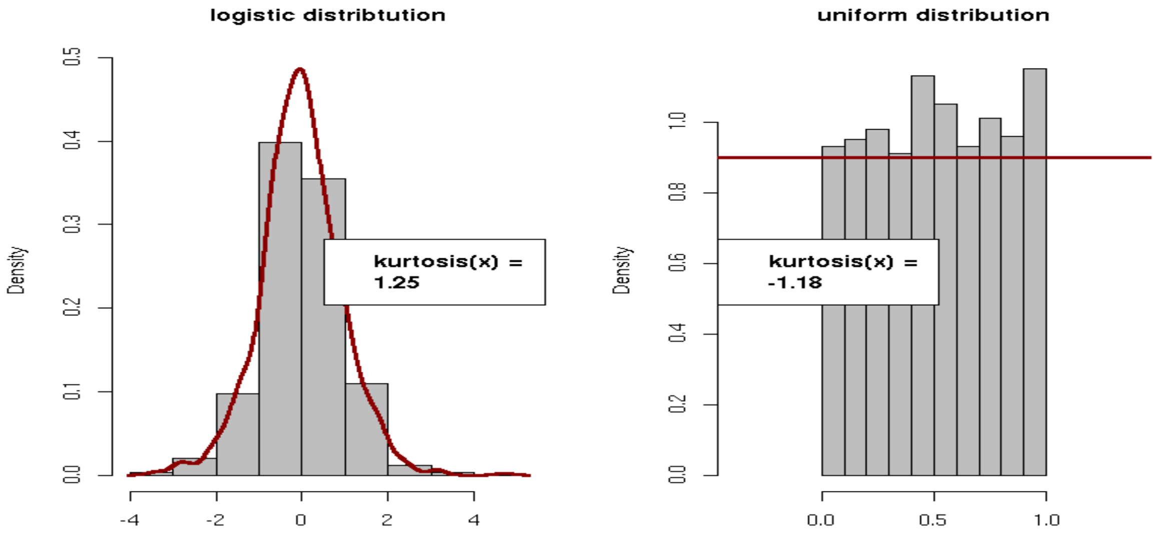


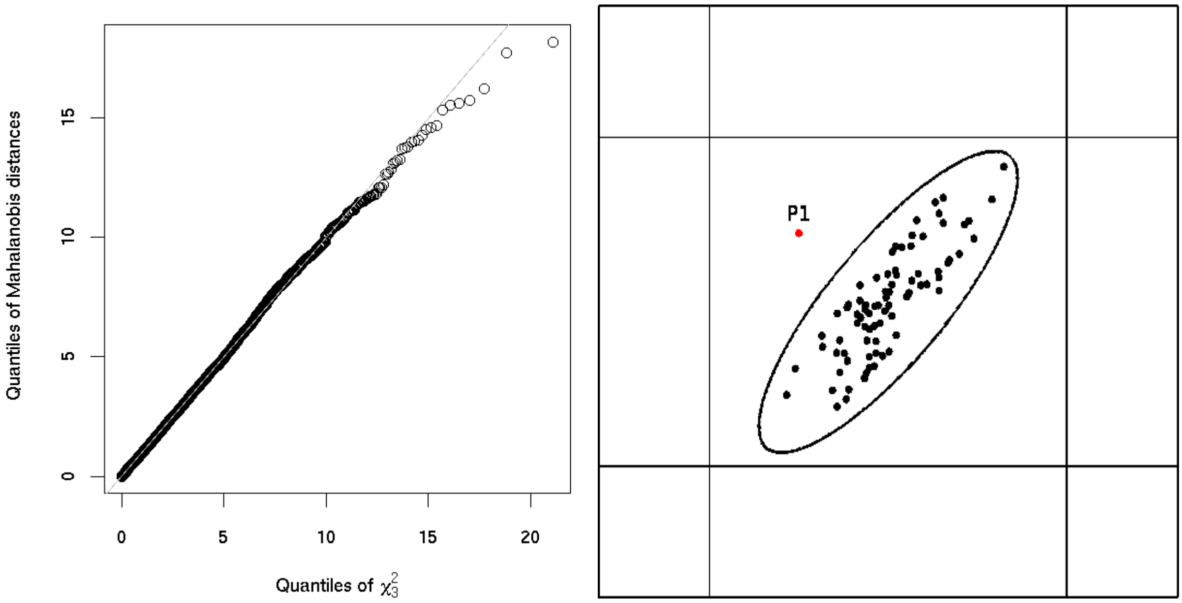
Figure 5.3: *The kurtosis measures whether a distribution is rather peak-shaped or flat relative to a Gaussian distribution.*

dimension p to a distribution follows a χ^2 distribution with p degrees of freedom (Mahalanobis, 1936). This allows us to define cutoffs for suspiciously large distances for a given confidence level α as in any statistical test. Figure 5.4 (a) exemplifies this: the figures gives the quantile-quantile plot (Q-Q plot) of the empirical quantiles of the Mahalanobis distances of 3000 random points in $\mathbb{R}^3$ and the quantiles of the χ^2 distribution with 3 degrees of freedom. The Q-Q plot is a diagnostic tool to determine if some data follow a specific distribution. We obtain a Q-Q plot by plotting the quantiles of a sample (in our case the Mahalanobis distances) against the quantiles of the distribution we want to test, in our case the χ^2 distribution. If all points lie on a straight line, the sample fits the assumed distribution well. The plot shows that the Mahalanobis distances fit the χ^2 distribution and that our assumption to use this distribution to determine significant outlier is valid.

Note that if we use this criterion for outlier detection, we effectively classify a map as outlier if its vector of quality descriptors differ by a large extent from the rest. This is reasonable since even for non-replicate LC-MS runs, we would expect most quality descriptors to be similar.

If we apply the Mahalanobis distance in this manner, we perform a multivariate outlier detection since we take all dimensions of the data at the same time into consideration. There is also the possibility of an univariate outlier detection e.g. testing each dimension independently. Figure 5.4 (b) illustrates this. The inner rectangle gives the confidence levels that would be defined by a univariate outlier method, which does only consider one dimension at a time. The outlier point P1 would not be classified as such by a univariate method. None of its individual dimensions deviate significantly from the rest of the points. In contrast to this, the ellipse illustrates the confidence levels of a multivariate outlier detection method. If applied, this method would correctly classify the point P1 as an outlier.

Nevertheless, our approach for outlier detection using the Mahalanobis distance suffers from two other drawbacks: first, the covariance matrix might be singular and not invertible, for instance if there are fewer LC-MS runs than descriptors ($n < p$). This can be seen from the fact the covariance matrix can be written as $\Sigma = \mathbb{E}(X^T X) - \mu^T \mu$. If $n < p$, X can have at most rank n and thus Σ, which has the dimensions $p \times p$, is singular. Second, outliers in the data might



(a) Q-Q plot of Mahalanobis distances and quantiles of the χ^2 distribution

(b) Multivariate vs. univariate outlier detection

Figure 5.4: *(a) Quantile-Quantile plot (Q-Q plot) of the empirical quantiles of the Mahalanobis distances of* 3000 *random points in* $\mathbb{R}^3$ *and the quantiles of the* χ^2 *distribution with* 3 *degrees of freedom. The plot shows that the Mahalanobis distance fit the* χ^2 *distribution well and that our assumption to use this distribution to determine significant outlier is valid. (b) Comparison between univariate (rectangle) and multivariate (ellipse) confidence intervals for outlier detection. Outlier point P1 would not be detected by a univariate method since a univariate method would only consider one dimension at a time. On the other hand, a multivariate outlier detection would classify this point correctly as outlier.*

distort our estimates of μ and Σ and lead to incorrect estimates of the distance. We solve the first problem by applying a principal component analysis to reduce the dimensionality of our data to a dimension $p' \ll p$ but try retain the essential information at the same time. We solve the second problem by using robust estimators for location and scale.

5.3.3 Principal Component Analysis

Principal component analysis (PCA) (Pearson, 1901) is a standard method for dimensionality reduction and feature extraction. Many textbooks and scientific articles cover its mathematical foundations and algorithmic components. In the following, we will introduce its key steps but for more details, we refer the interested reader to the literature (Smith, 2002; Wold *et al.*, 1987).

By performing PCA, our aim is to represent a data vector x by a lower dimensional representation given by $Mx = y$. M is a matrix with dimension $dim(y) \times dim(x)$ with $dim(y) < dim(x)$. M represents a projection from the higher dimensional space of x to a lower dimensional space of y. In the case of PCA, M is an orthogonal linear projection. We implicitly assume that our high-dimensional data x lies close to a hyperplane and that we can approximate each data point using the vectors that span this hyperplane. Mathematically, this is expressed as:

$$x \approx m + \sum_{i=1}^{n} w_i b_i \qquad (5.17)$$

We choose the vectors $b_i, i = 1, \ldots, n$ to be orthonormal, that is $(b_i)^T b_j = 0$ for $i \neq j$ and

$(b_i)^T b_j = 1$ for $i = j$. Obviously, we hope to approximate our data well with a small number of vectors n, much smaller than the dimension of the data. The standard PCA consists of the following steps:

1. The data, in our case the vectors of quality descriptors for each map, are stored in a $n \times p$ matrix X with a row for each of the n maps and a column for each of the p quality descriptors. This matrix is centered by subtracting the column-wise mean m.

2. Compute the covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ of the data. It contains the variance of each dimension p on its diagonal and the covariances in the remaining entries, as we explained in Sect. 5.2. Compute the eigenvectors of the covariance matrix and choose as b_i the eigenvectors with the largest eigenvalues.

3. Project each data vector x into the lower dimensional space spanned by the chosen principal components by computing $y = Ex$ where E contains the chosen eigenvectors as column vectors.

4. We can approximately reconstruct each data by computing:

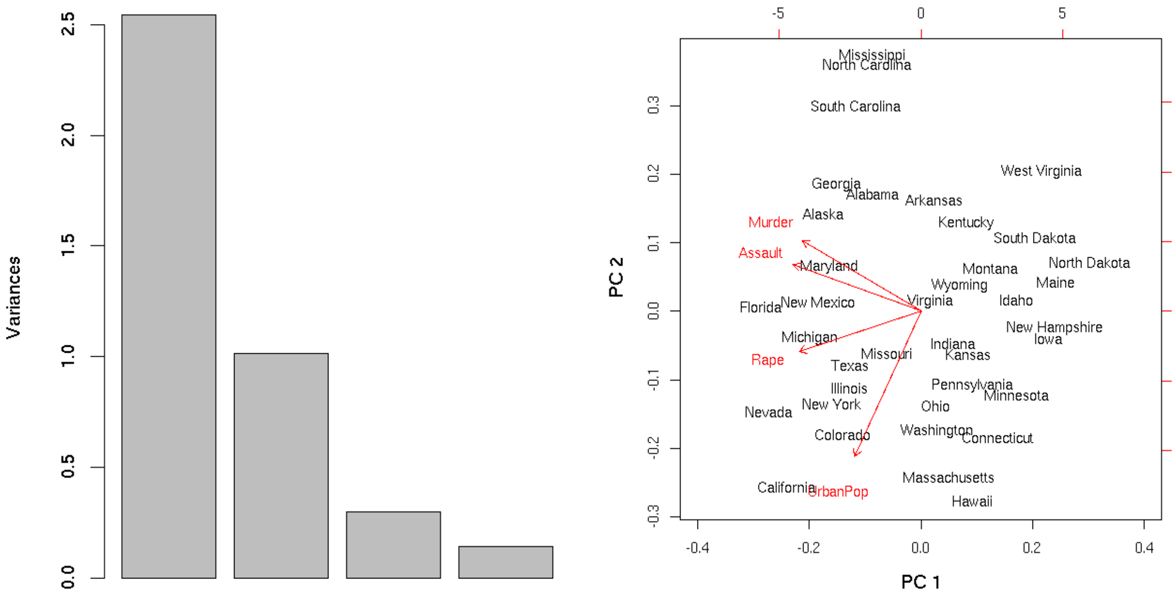
$$x \approx m + Ey \tag{5.18}$$

The magnitude of each eigenvalue represents the amount of variance that the corresponding eigenvector captures. We illustrate this fact by a simple example. Assume that the dimensions of our data are pairwise uncorrelated. In this case, the covariance matrix is zero everywhere except on its diagonal where each entry corresponds to the variance of the corresponding dimension:

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & 0 & \dots \\ 0 & \sigma_{22} & 0 & \dots \\ \dots & & & \\ \dots & & 0 & \sigma_{pp} \end{pmatrix} \tag{5.19}$$

The first eigenvector of this matrix is $v_1 = (1, 0, 0 \dots 0)$ and the eigenvalue σ_{11} which is the variance of the first dimension, the second eigenvector is $v_2 = (0, 1, 0 \dots 0)$ with eigenvalue σ_{22} and so on. Thus, if we sort the eigenvectors by the magnitude of their eigenvalues, choose the eigenvectors with the largest eigenvalues and project our data into the space spanned by these vectors, we have chosen new coordinates or principal components for our data which represent the highest variance. The entries of the eigenvectors are called *loadings* in PCA terminology. There are usually interpreted as weightings of the original variables. We refer to the projection of the data on the individual principal components as *scores.*

It is not straightforward to decide on the number of principal components to use. A useful tool is the so-called *scree plot.* This is simply a bar chart of the principal components, ordered by the magnitude of their eigenvalue. The height of each bar is given by this magnitude. Ideally, we would like to see a small number of principal components that together capture most of the variance. Figure 5.5 shows an example where the first and second principal component capture a significant amount of the variance whereas the remaining two contribute much less. The plot was created using the example data set *USarrests*, included in the R software package (http://www.r-project.org). This data set contains the arrests per 100000 residents for assault, murder, and rape in each of the 50 US states in the year 1973. It also contains the percent of the population living in urban areas for each state. Given the scree plot, it seems sensible to reduce



(a) Scree plot

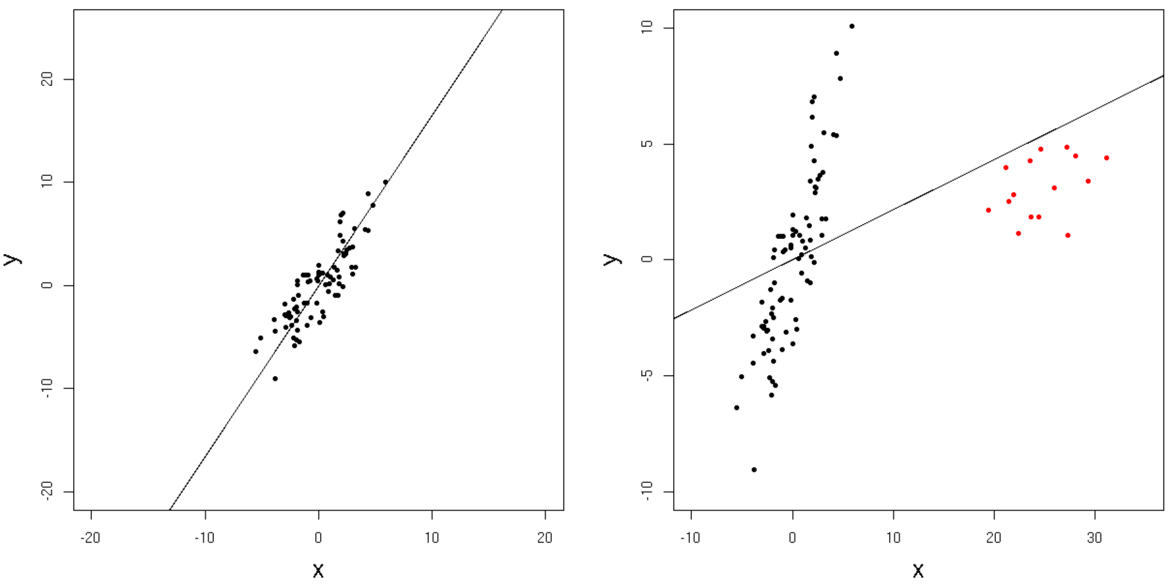
(b) Biplot of first two principal components

Figure 5.5: *(a) The scree plot for a principle component analysis. It is a bar chart of the principal components and their variances. The height of each bar gives the variance of each component. The data is the example data set USarrests, included in the R software for statistical computing (http://www.r-project.org). We can see that the first two principal components capture most of the variance* (86.75%) *whereas the third and the fourth contribute only little information. In this case, it would make sense to project the data on the subspace spanned by the first two principal components. (b) The biplot for the same data. It reveals that the first component differentiates between US states with low and high crime rates in all three categories whereas the second component differentiates between states with high and low urban population. We do not show all states to keep the plot readable.*

this four-dimensional data set to the two-dimensional space spanned by the first two principal components and to discard the remaining components since the first two principal components capture 86.75% of the variance.

The second diagnostic plot that is frequently used in conjunction with PCA is the *biplot*. The biplot is a plot of the data projected onto its first two principle components. It also contains a plot of the variables which are represented as arrows in the space of the first two principle components. Figure 5.5 (b) shows a biplot for the *USarrests* data set. It reveals that the first component differentiates between US states with low and high crime rates in all three categories whereas the second component differentiates between states with high and low urban population.

A problem with the standard approach to PCA is that it is strongly influenced by outliers in the data. In statistical terms, outliers are samples with an 'extreme characteristic due to at least one atypical value of the measured parameters' (Daszykowski *et al.*, 2007). This is illustrated by Figure 5.6. The left plot shows a two-dimensional data set with its first principal component. The right plot shows the same data but with several outlier points colored in red. As we can see, the outliers severely influence our estimate of the principal component. The next section introduces a robust version of PCA, which offers a solution to this problem.



(a) First principal component

(b) First principal component in the presence of outliers

Figure 5.6: *Influence of outlier points on the principal components. (a) A set of two-dimensional points with their first principal component. (b) The same data with a small group of outlier points (red). Note the influence of the outliers on the principal component.*

5.3.4 Robust Principal Component Analysis

As we described above, the standard PCA approach is sensitive to outliers. Outlier points will lead to wrong estimates of center and scale and thus distort the results of the projection. To remedy this, we use a robust version of principal component analysis (rPCA). In statistics, an approach is considered robust if it is not, or not severely, influenced by outlier observations. We use a rPCA algorithm developed by Croux and Ruiz-Gazen (1996). Recall that the standard PCA algorithm centers the data by subtracting the mean and computes the eigenvectors of the covariance matrix. Both, estimates of the mean and covariance matrix are severely affected by outlier points. A robust approach to PCA must therefore consist of robust estimate of the mean (location) and a robust estimation of the covariance matrix (scale or spread).

An important concept in robust statistics is the *breakdown point* of an estimator. The breakdown point of an estimator is the fraction of data that can be given arbitrary values without making the estimator arbitrarily bad. The higher the breakdown point, the more robust the estimator is. For example, the sample mean has a breakdown point of 0, because we can make it arbitrarily large just by changing any of the sample components. This is clearly not desirable and thus the mean is not a good estimator of location in the presence of outlying observations.

Intuitively, we can see that a breakdown point cannot exceed 0.5 since if more than half of the observations are contaminated, we cannot distinguish between the underlying distribution and the outlier distribution. Therefore, the maximum breakdown point is 0.5. There are estimators, such as the median, that have the maximum breakdown point.

The next two sections introduce two algorithms that address the problems of standard PCA: a robust estimator of location, the L_1 median, and an approach to estimate a robust replacement of the principal components, e.g. the eigenvectors of the covariance matrix.

A steepest Descent Algorithm to compute the L_1 Median

The L_1 median is a robust estimator of location and is defined as:

$$\mu_{L1} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{m} \|x_i - \theta\|_2 \tag{5.20}$$

where $\| * \|_2$ denotes the Euclidean distance in $\mathbb{R}^n$. The L_1 median is simply the point θ, not necessarily a data point, which minimizes the Euclidean distance to all other points $x_i, i = 1, \ldots, m$. It is also known as the *geometric median* or *Fermat-Weber point.* The problem of computing the L_1 median also has an idealized interpretation as the optimal selection of a location. As an example, if a company wants to choose the location of a warehouse which has to serve m customers, the warehouse should be placed at a location such that the distance to the customers is minimal.

In the univariate case, the L_1 median is equal to the standard median. For multivariate data, the L_1 median gives a robust estimate of location in contrast to the mean or component-wise median. Furthermore, it is invariant to orthogonal linear transformations such as PCA and has the maximum breakdown point of 0.5 (Lopuhaa and Rousseeuw, 1991) and is thus highly robust. There exist several algorithms to compute it efficiently (Bose *et al.*, 2002; Vardi and Zhang, 2000; Weiszfeld, 1937). We decided to use an approach based on steepest descent with stephalving (Hössjer and Croux, 1995). An implementation already exists in the R package *pcaPP*, which is not the case for all competing approaches. To iteratively compute an estimate $\hat{\mu}$ of the L_1 median using the pcaPP algorithm, we define the objective function S as:

$$S(\hat{\mu_n}) = D_n(\|x_i - \mu\|, \ldots, \|x_n - \mu\|)$$

where D_n is a rank-based distance measure (Hössjer and Croux, 1995) defined as:

$$D_n(v) = \sum_{i=1}^{n} R(|v_i|)|v_i|$$

where $R(|v_i|)$ is the rank of $|v_i|$ among $|v_1|, \ldots, |v_n|$. The algorithm uses a weighting function $w_i(\mu)$ defined as:

$$w_i(\mu) = \begin{cases} \frac{R(||x_i - \mu||)}{||x_i - \mu||} & \text{if} \quad x_i \neq \mu \\ 0 & \quad\quad x_i = \mu \end{cases} \tag{5.21}$$

where $\| * \|$ is the Euclidean distance, $R(x)$ is the rank of x. This function assigns a higher weight to sample points with a large distance to the current estimate of the L_1 median, μ. Furthermore, Hössjer and Croux (1995) use a gradient vector

$$\Delta(\mu) = \frac{\sum_{i=1}^{n} w_i(\mu)(x_i - \mu)}{\sum_{i=1}^{n} w_i(\mu)}. \tag{5.22}$$

The solution of the minimization problem satisfies $\Delta(\mu) = 0$ except when we have a degenerate solution. Most other algorithms use similar objective functions and an iterative procedure to approximate to the L_1 median in a stepwise manner.

Our steepest descent algorithm is given in Algorithm 1 as pseudo-code. From a first starting solution $\hat{\mu}_{n,k}$, we compute the gradient $\Delta(\hat{\mu}_{n,k})$. As an example, the coordinate-wise mean will give a reasonable starting point for the optimization. We start the steepest descent by taking a step $\Delta(\hat{\mu}_{n,k})$ from $\hat{\mu}_{n,k}$, if this leads to a decrease of the objective function S. If not, we reduce the

```
Initialize $tol$ and $maxstep$ to control the precision;
Set k=0 and $\hat{\mu}_{n,0}$ to an initial estimator, for example the coordinate-wise median
while $k \leq maxstep$ do
    if $||\Delta(\hat{\mu}_{n,k})|| < tol$ then
        $maxhalf = 0$
    else
        $maxhalf$ = (ln($\|\Delta(\hat{\mu}_{n,k})\|$) - ln(tol))/ln(2)
    end
    $\hat{\mu}_{n,k+1} = \hat{\mu}_{n,k} + \Delta(\hat{\mu}_{n,k})$;
    $j = 0$;
    while $S(\hat{\mu}_{n,k+1}) > S(\hat{\mu}_{n,k})$ and $j \leq maxhalf$ do
        j = j+1;
        $\hat{\mu}_{n,k+1} = \hat{\mu}_{n,k} + \Delta(\hat{\mu}_{n,k}/2^j)$;
    end
    if $j > maxhalf$ then
        Return ($\hat{\mu}_{n,k}$);
    end
    k = k+1;
end
```

Algorithm 1: The steepest descent algorithm to compute the L_1 median (Hössjer and Croux, 1995).

step size to $2^{-j}\Delta(\hat{\mu}_{n,k})$ where j is the smallest integer i such that $S(\hat{\mu}_{n,k}+2^{-i}\Delta(\hat{\mu}_{n,k})) < S(\hat{\mu}_{n,k})$. After this, we set the estimate $\hat{\mu}_{n,k+1}$ to $\hat{\mu}_{n,k}+2^{-j}\Delta(\hat{\mu}_{n,k})$. In this way, the value of the objective function decreases with every step. We stop the iteration when the objective function does not decrease after a step smaller than the variable *tol* in the direction of $\Delta(\hat{\mu}_{n,k})$. A second control variable is *maxstep*, which gives an upper bound on the number of optimization steps. Default values of *maxstep* are 200 and 10^{-8} for *tol*.

According to Hössjer and Croux (1995), the combination of a rank-based weighting function and the step-halving during the steepest descent should guarantee convergence. Their experiments confirmed this but they were not able to prove it in a formal manner. In our experience, this algorithm is also quite fast: computing the L_1 median for our data took less than $1s$ (tests performed on 3.2 GHz Intel Xeon CPU with 3 GB memory running Debian Sarge).

A Projection Pursuit Approach to Principal Component Analysis

As we already mentioned above, the standard principal component analysis is very sensitive to outlier observations. This has been pointed out by several authors and has led to several robust versions of PCA (Daszykowski *et al.*, 2007). One approach consists in calculating robust estimates of eigenvalues and eigenvectors, without explicitly computing the covariance matrix. This method is referred to as robust PCA using projection pursuit (Croux and Ruiz-Gazen, 1996). Projection pursuit is a general statistical technique: we aim at finding the most interesting projections in a multivariate data set (Friedman and Tukey, 1974). Usually, 'interesting' is defined in terms of deviation from a Gaussian distribution or having a high variance.

In short, the projection pursuit approach to robust PCA examines a finite set of candidate directions in the measurement space. In our case, the set of candidate directions is given by the centered observations themselves. We fix the first direction which maximizes a robust estimator of scale and choose all following direction vectors to be orthogonal to the previous one. We use the *Median of Absolute Deviation* (MAD) to estimate the scale :

$$\mathrm{MAD} = \mathrm{median}_i(|x_i - \mathrm{median}_j(x_j)|). \tag{5.23}$$

In other terms, we obtain the MAD by computing the absolute deviations (or residuals) from the median and compute their median. The MAD is a more robust measure of spread as, for instance, the variance. This is due to the fact that the distances to the mean in the variance are squared, so large distances which might be caused by outliers, will have a lot of influence on the estimate.

The definition of the MAD as given in Eq. 5.23 already gives us a robust measure of spread or scatter. But if we want to use the MAD as a consistent estimator of the standard deviation, we need to apply a correction factor c which results in $\sigma_{\mathrm{MAD}} = c \times \mathrm{MAD}$. For Gaussian distributed data, c is taken to be $1/\Phi^{-1}(3/4) \approx 1.4826$, where Φ^{-1} is the inverse of the quantile function of the Gaussian distribution with mean 0 and standard deviation 1. This is because the MAD is given by:

$$\frac{1}{2} = P(|X - \mu| \leq \mathrm{MAD}) = P\left(\left|\frac{X-\mu}{\sigma}\right| \leq \frac{\mathrm{MAD}}{\sigma}\right) = P\left(|Z| \leq \frac{\mathrm{MAD}}{\sigma}\right) \tag{5.24}$$

This means that, for any symmetric distribution, the MAD is the distance between the first and second quartiles and thus the MAD equals the 75th percentile. Since we assume Gaussian distributed data, the scale factor c is the 75th percentile of the Gaussian distribution with $\sigma = 1$. This results in

$$\frac{\mathrm{MAD}}{\sigma} = \Phi^{-1}(3/4) \approx 0.6745. \tag{5.25}$$

and

$$\sigma_{\mathrm{MAD}} \approx 1.4826\ \mathrm{MAD}. \tag{5.26}$$

The MAD has a breakdown point of 50%, which is the highest possible value (Lopuhaa and Rousseeuw, 1991).

After projecting the data into the subspace of the direction with the highest robust variance, projection pursuit searches for an approximation of the next eigenvector which is assumed to be orthogonal to the previous one. The pseudo-code description of the projection pursuit approach to PCA (Croux and Ruiz-Gazen, 1996) is given in Listing 2.

In this context, the Median of Absolute Deviation is called the *projection index*. In the for-loop starting at line 3, we enforce that all data vectors are orthogonal to the previously computed eigenvector $\hat{v}$ by first projecting the data vector on this eigenvector (the result of this projection is called *score*, as we explained in Sect. 5.3.3). We proceed by standardizing the data vectors and choose the next principal component as the direction which maximizes the projection index in line 8. This procedure is illustrated in Fig. 5.7. Note that this algorithm bears some similarity to the Gram-Schmidt process which is an algorithm to orthogonalize a finite set of linearly independent vectors.

Obviously, one needs to decide on the dimensionality of the subspace the data is projected in. We always choose a number of components that would explain 90% of the variance, which is usually around 6. Consequently, we can now describe each LC-MS map using a vector in 6 dimensions instead of 20, the covariance matrix Σ is invertible and we can search for suspicious maps by plotting the Mahalanobis distance for each map. Note that each dimension does not longer represent an unique descriptor, but a linear combination of all descriptors. But by

```
input  : A multivariate data set $X = \{x_1^1, x_2^1, \ldots, x_n^1\}$, $q =$ desired number of principal
         components
output: The robust eigenvectors $\{\hat{v}_{\text{MAD},1}, \ldots, \hat{v}_{\text{MAD},k}\}$
Center the data by subtracting the $L_1$ median;
Obtain the first candidate direction by

    $A_{n,1}(X) = \left\{ \frac{x_i^1}{||x_i^1||}; 1 \leq i \leq n \right\}$

and compute the first principal component as

    $\hat{v}_{\text{MAD},1} = \arg \max_{a \in A_{n,1}(X)} \text{MAD}(a^t x_1^1, \ldots, a^t x_n^1)$

Compute the scores on the first component as $y_i^1 = \hat{v}_{\text{MAD},1}^t x_i^1$ for $1, \ldots, n$;
for $k = 2, \ldots, q$ do
    for $i = 1, \ldots, n$ do
        $x_i^k = x_i^{k-1} - y_i^{k-1} \hat{v}_{\text{MAD},k-1}$;
    end
    Define the set $A_{n,k}(X) = \{\frac{x_i^k}{||x_i^k||}; 1 \leq i \leq n\}$;
    Define the estimated eigenvector $\hat{v}_{\text{MAD},k} = \text{argmax}_{a \in A_{n,k}(X)} \text{MAD}(a^t x_1^k, \ldots, a^t x_n^k)$;
    for $i$=1,...,$n$ do
        $y_i^k = \hat{v}_{\text{MAD},k}^t x_i^k$;
    end
end
```

Algorithm 2: Robust PCA: a projection pursuit approach (Croux and Ruiz-Gazen, 1996).

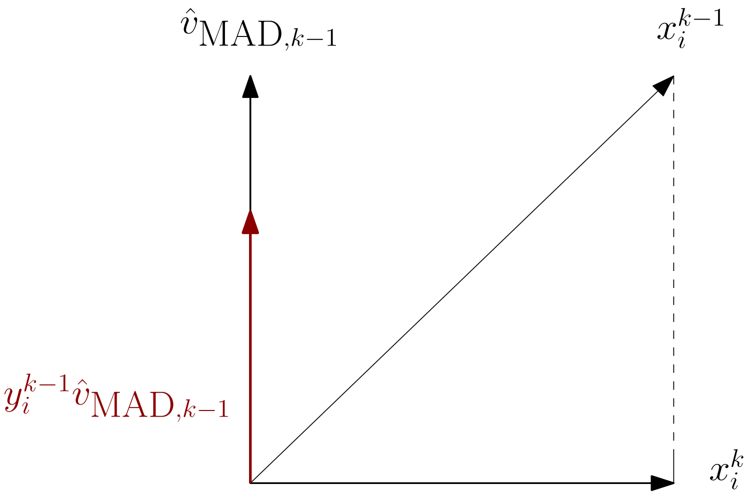

$\hat{v}_{\text{MAD},k-1}$
x_i^{k-1}
$y_i^{k-1}\hat{v}_{\text{MAD},k-1}$
x_i^k

Figure 5.7: *Illustration of the projection pursuit approach: x_i^{k-1} is projected on $\hat{v}_{MAD,k-1}$. The resulting vector is subtracted from x_i^{k-1} to obtain x_i^k which is orthogonal to $\hat{v}_{MAD,k-1}$. This is repeated for all data vectors x_i^{k-1} to obtain a new set of candidate directions which are all orthogonal to the previously computed principal component $\hat{v}_{MAD,k-1}$.*

inspecting the weights (also called *loadings* in PCA terms) we can gain insights into which descriptor contributed the most to a particular dimension.

Using this projection pursuit algorithm, we have now obtained a robust dimension reduction of our data. This solves the problem that the covariance matrix is singular if $n < p$ i.e. if we have less samples than descriptors. But we also need a robust estimate of the covariance matrix to compute the Mahalanobis distance. Note that the kth eigenvalue is by definition the robust variance of the data projected on the kth eigenvector:

$$\lambda_{\text{MAD},k} = \text{MAD}^2((v_{\text{MAD},k})^t x_1, \ldots, (v_{\text{MAD},k})^t x_n)$$

Using the eigenvalues, we can construct a robust estimate of the covariance matrix:

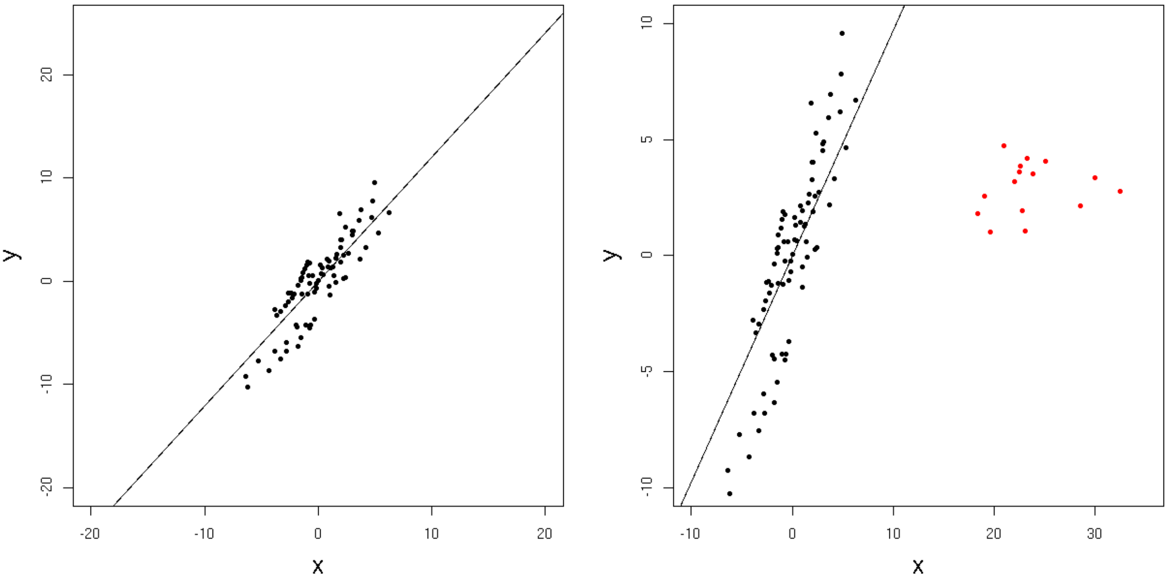
$$\Sigma_{\text{MAD}} = \sum_{k=1}^{p} \lambda_{\text{MAD},k} \cdot v_{\text{MAD},k} \cdot v_{\text{MAD},k}^t$$

According to Li and Chen (1985), the advantage of this approach is that these estimates inherit the breakdown point of the scale estimator MAD which is 0.5, the highest possible value. Figure 5.8 illustrates the robustness of the projection pursuit approach: the first principal component is barely affected by the outlier points.

To sum up, this projection pursuit approach to PCA has helped us to solve two problems: the new estimate of the covariance matrix is not longer singular and thus can be inverted to compute the Mahalanobis distance. Second, we used robust estimates of scale and location to limit the influence of outliers. In the following sections, we will apply our method to several data sets and demonstrate its versatility.

5.4 Results

We present three use cases to demonstrate how our approach can be applied to automatically detect outlier runs among a set of LC-MS maps. We start with a set of simulated maps. The simulation allows us to probe the capabilities of our approach on a detailed level. The second



(a) First robust principal component

(b) First robust principal component in the presence of outliers

Figure 5.8: *Influence of outlier points on robust principal components. (a) A set of two-dimensional points with their first principal component. (b) The same data with a small group of outlier points (red). Note that, in contrast to the standard PC estimation, the influence of the outliers on the principal component is negligible.*

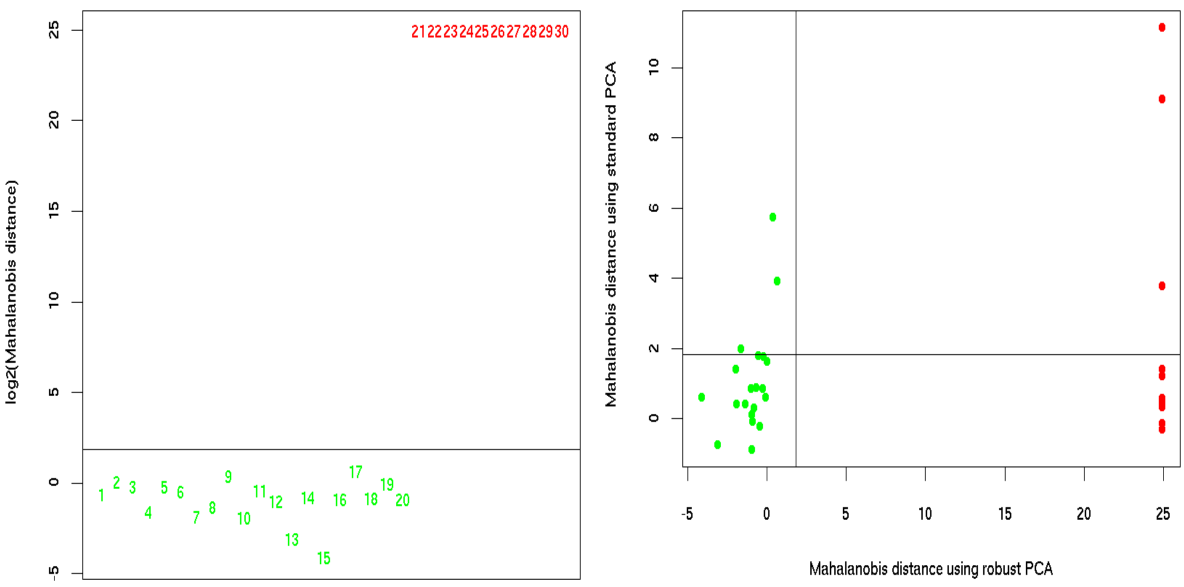
and third use case comprise a tryptic digest of bovine serum albumine (BSA) and urine samples from a healthy volunteer recorded using LC-ESI-MS.

Details of the full mass spectrometry analysis concept and chromatographic conditions are described elsewhere (Machtejevas *et al.*, 2007). In short, we employed a restricted access sulphonic acid strong cation-exchanger (RAM - SCX) (Merck KGaA, Darmstadt, Germany) column followed by a peptide transfer and solvent switch thought trap column (Chromolith Guard, 5×4.6mm Merck KGaA). We performed a subsequent analysis using an analytical column (Chromolith CapRod RP18e, 150 × 0.2mm, Merck KGaA) by means of column switching to perform two dimensional orthogonal separations. On-line mass spectrometric detection was performed using an Esquire Series 3000 PLUS ESI ion trap mass spectrometer (Bruker Daltonics, Bremen, Germany).

The BSA digest was prepared according to a standard procedure (Proteo Extract All-in-one Trypsin Digest Kit, Merck Chemicals Ltd, Nottingham, UK) with final concentration of 2 mg/ml and stored at 20° C. Urine samples were from healthy volunteers pooled and stored at 20° C. Before analysis samples were defrosted at room temperature for an hour, and then filtered through 0.22μm pore size low protein binding membrane filters (Durapore, Millipore) and clear sample transferred to autosampler tubes. The prepared samples were stored in an autosampler at 4° C not longer than 24h before injection.

5.4.1 Simulated LC-MS Runs

To provide a sanity check of our method, we simulated a mixture of standard peptides using our software LC-MSsim (Schulz-Trieglaff *et al.*, 2008a). This software simulates an entire LC-MS experiment, including protein digestion, retention time prediction, isotope pattern and elution



(a) Mahalanobis distances

(b) Robust vs. non-robust distances

Figure 5.9: *(a) Log-scaled Mahalanobis distances for the simulated LC-MS runs. The black line gives a cutoff for a significance level of* 5% *as computed from the* χ^2 *distribution. (b) Examples of LC-MS maps classified as good (green) and outlier (red).*

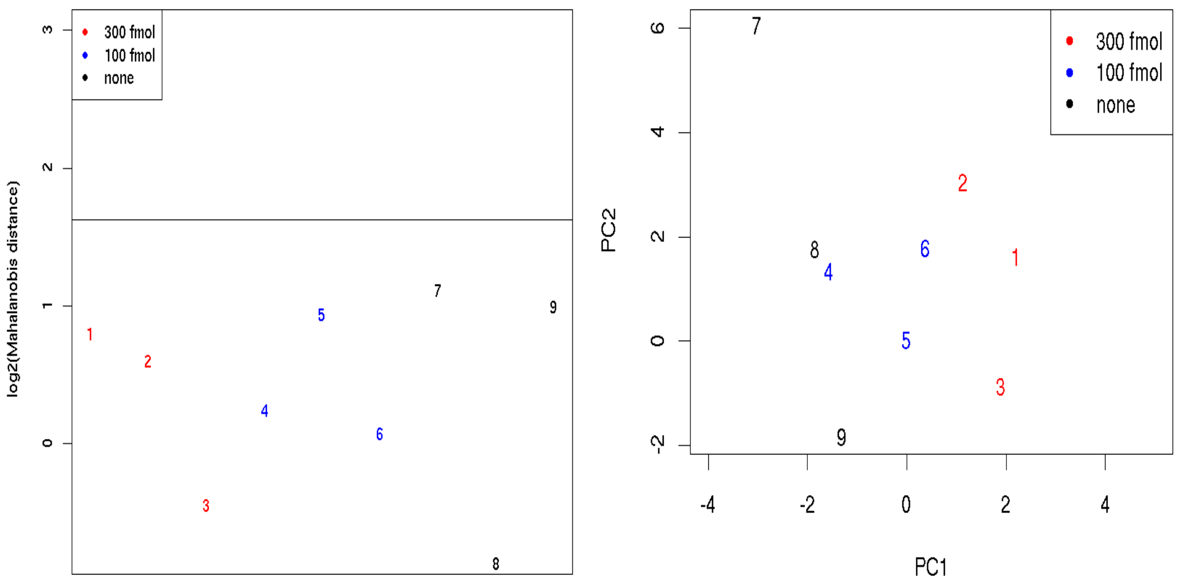
peak models. It produces a realistic LC-MS map and the user can introduce noise, non-peptidic contaminants, m/z and intensity errors at its own will. We provided details of this software in the previous chapter.

We simulated a mixture consisting of peptides from a tryptic digest of 4 standard proteins, namely bovine catalase, horse myoglobin, bovine carbonic anhydrase and bovine lysozyme. To test our approach, we simulated ten perfectly reproducible runs and three outliers. In the first simulated outlier, all peptides elute with a high variance of retention time. The second contains many spectra with a significantly larger amount of shot noise whereas the third contains only very noisy elution peaks.

Our aim was to investigate whether our approach based on quality descriptors, Mahalanobis distance and robust principal component analysis would be able to recover all outlier runs correctly. Figure 5.9 (a) shows that the robust Mahalanobis distance highlights all the simulated outlier with a significantly large distance. Furthermore, Fig. 5.9(b) shows a comparison of the robust Mahalanobis distance versus a non-robust version of this distance (i.e. without robust PCA and robust estimator of location). The black lines indicate the cutoff for a confidence level of $\alpha = 0.05$ with the Bonferroni correction for multiple testing. This plot shows that it actually makes sense to use robust PCA since the standard PCA would incorrectly classify several good maps as outlier and not detect several true outlier.

5.4.2 Influence of Differential Expression

The question arises whether our method is influenced by LC-MS maps with signals of differentially expressed peptides. These differential signals will necessarily lead to differences in the ion count of several spectra and might even influence the shape of the TIC. But if these are the only differences and if these differences are not too strong, our method should not classify the LC-MS maps as outlier. It is not immediately clear whether our method is robust enough for



(a) Mahalanobis distances

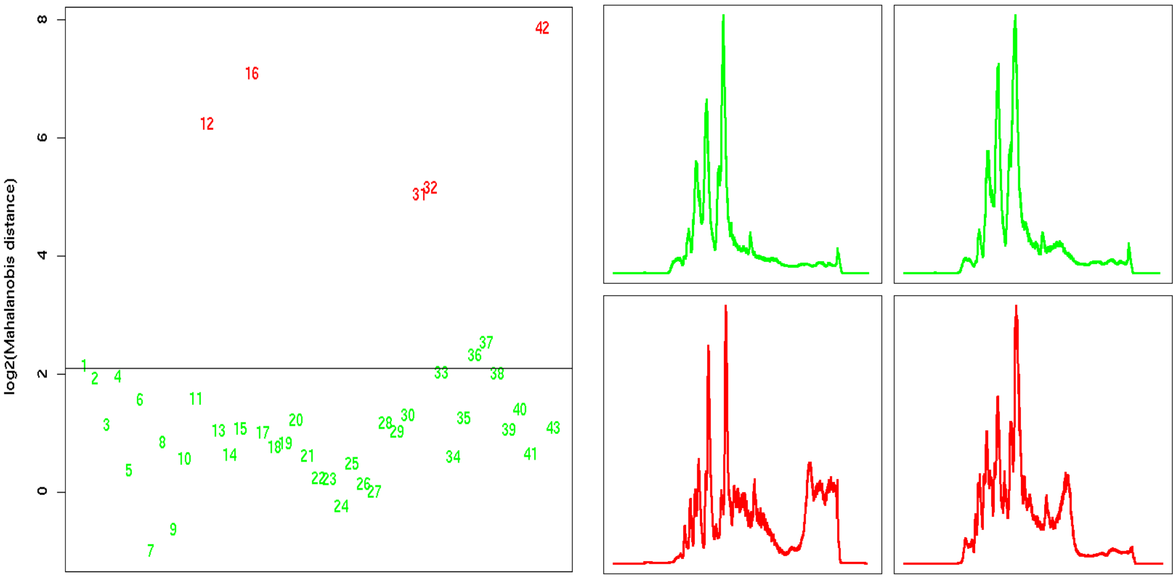
(b) Projection on the first two principal components

Figure 5.10: *(a) Log-scaled Mahalanobis distances for the* 9 *LC-MS runs with spiked-in peptides. The black line gives a cutoff for a significance level of* 5% *as computed from the* χ^2 *distribution. LC-MS maps spiked with* 300 *fmol BSA are colored in red, blue indicates* 100 *fmol and black no spiked-in BSA. We can see that all LC-MS maps, also the ones with 'artificial differential expression', are far below the threshold for outlier. (b) Projection of the LC-MS maps on the first two principal components. This plot does reveal a clustering of the maps according to the concentration spiked-in, but it does not influence the outlier detection.*

this purpose.

To investigate this matter, we obtained a set of 9 LC-MS maps recorded from a digest of HeLa cell samples. This immortal cell line was derived from human cervical cancer cells. The cells were originally taken from Ms. Henrietta Lacks and HeLa is the abbreviation of her name. In short, fifty micrograms of HeLa cell lysate protein was solubilized in 6 M urea, 2 M thiourea, 25 mM Tris (pH 8.0), reduced (1 ug dithiothreitol, 30 min), alkylated (5 ug iodoacetamide, 30 min) and cleaved to peptides (1 ug LysC, 3 h, then diluted 4x with 50 mM ammonium bicarbonate, then 1 ug trypsin, overnight) (REF 16615899). We withdrew 9 aliquots and spiked three of them with 300 fmol of previously-digested bovine serum albumin (BSA). Three other aliquots were spiked with 100 fmol BSA whereas the remaining 3 samples were left unchanged. The spiked-in BSA peptides constitute our artificially created differential expression. All aliquots were cleaned up using STAGE tips (REF 12585499) and analyzed by LC-MS/MS on an Agilent 1100 Series nanoflow HPLC coupled to an LTQ-Orbitrap. The LC-MS maps were supplied by Dr. Mark Robinson at the Walter and Eliza Hall Institute (WEHI) in Melbourne, Australia.

Our aim was to verify whether differential expression would distort the results of our computational quality assessment. All LC-MS maps are of high quality and exhibit an excellent reproducibility in terms of retention time and intensities and therefore none of them should not be declared as outlier. Figure 5.10 (a) shows the Mahalanobis distances of the 9 LC-MS maps estimated using our robust algorithm. The black horizontal gives the cutoff based on the χ^2 for a confidence level of $\alpha = 0.05$. LC-MS maps spiked with 300 fmol BSA are colored in red, whereas blue indicates 100 fmol and black no spiked-in BSA. We can see that all maps are well below this threshold and are thus not identified as outliers. Interestingly, the projection on the



(a) Mahalanobis distances

(b) Total ion chromatograms of good (top) and outlier maps (bottom)

Figure 5.11: *(a) Log-scaled Mahalanobis distances for the BSA digest. The numbers denote the time order of the runs. The black line gives a cutoff for a significance level of* 5%. *(b) Examples of LC-MS maps classified as good (top row) and outlier (bottom row).*

first two principal components, shown in Fig. 5.10 (b), does reveal a clustering of the LC-MS maps according to the spike-in concentration. This does not affect the outlier detection which we performed on 6 principal components which captured 91.5% of the variance.

5.4.3 Tryptic Digest of Bovine Serum Albumin

This data set consists of replicate LC-MS recordings of a tryptic digest of bovine serum albumin (BSA). The peptide mixture was measured in 43 replicates, details of sample preparation and LC-MS analytics are described elsewhere (Machtejevas *et al.*, 2007). Using the algorithm implemented in TOPP/OpenMS (Kohlbacher *et al.*, 2007; Sturm *et al.*, 2008), we performed peptide feature detection, alignment and statistical analysis for these runs. After manual inspection, we classified 5 runs as outliers for various reasons: 3 exhibited peptide feature intensities that deviated by a large extent from the other replicates. The remaining two revealed significant shifts in retention time as compared to the remaining runs. This fact made an alignment difficult and required manual fine-tuning of the alignment algorithm.

This is of course a time-consuming procedure. It would be preferable to have a method that would allow us to remove outliers before feature detection and alignment is performed to save time and computer resources. Consequently, we applied our quality assessment method to these runs.

Figure 5.11 (a) shows for each of the 43 replicates the Mahalanobis distance D_M to the center of all other measurements. LC-MS maps that were manually identified as outlier are colored in red, good replicates in green. The horizontal line gives a cutoff corresponding to a significance test with $\alpha = 0.05$ and Bonferroni correction for multiple testing. In other terms, all LC-MS maps with distances above this threshold are classified as outlier by our method.

As we can see from Fig. 5.11 (a), the combination of spectral quality descriptors, robust principal component analysis and Mahalanobis distance is accurate and classifies all outlier maps correctly. It also classifies some additional maps as mild outliers, namely the first LC-MS map and the maps with number 36 and 37. Manual inspection of the PCA loadings revealed that their larger Mahalanobis distances are mainly due to a higher noise level in some spectra and minor fluctuations in the TIC. For illustration, Fig. 5.11 (b) shows the TIC of four maps. Again, normal runs in the upper row are colored in green, outlier runs are colored in red. Both outliers maps exhibit TICs that contain a significant amount of noise peaks and clearly deviate from the two good runs in the top row. Whereas the left outlier map can easily be identified from its TIC, the second outlier is less pronounced and might escape a superficial visual inspection.

Another interesting observation is the fact that even the LC-MS maps not classified as outliers seem to fall into two different classes. Maps with an even number seem to cluster together whereas runs with odd numbers fall into a different group. This can be explained by the fact that we employed two chromatographic columns during this experiment: the first BSA replicate was analyzed using the first column, the second replicate with the second column, the third again with the first column. We used this setup to achieve a higher sample throughput. It is nice that our method visualizes this particular characteristic of our setup and proves that our method is able to recover subtle tendencies in the data.

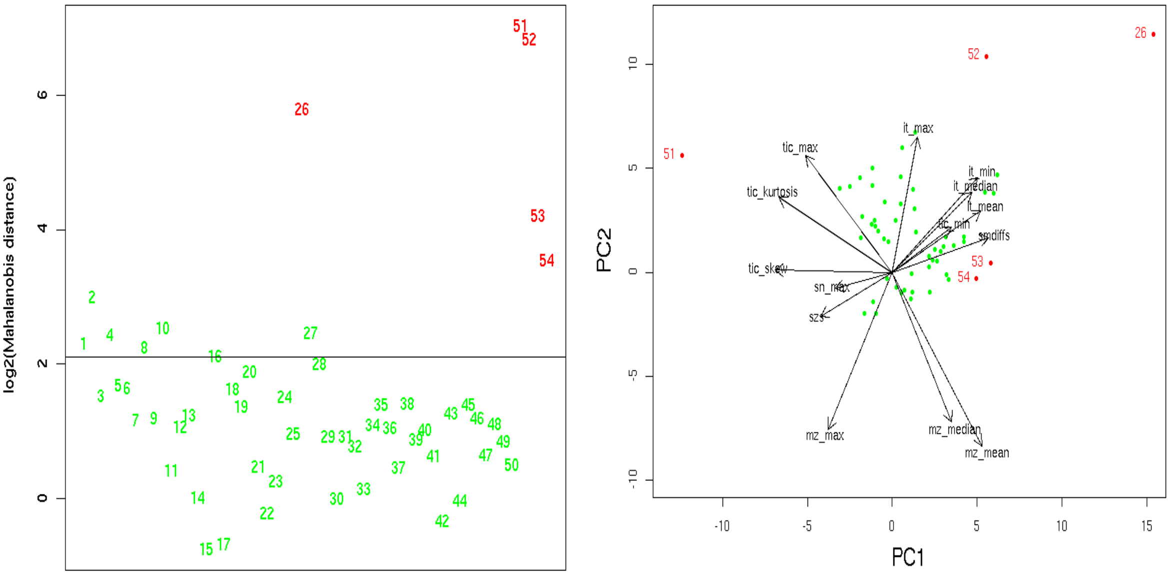
5.4.4 Urine Samples of a Healthy Volunteer

This data set consists of 54 LC-MS runs. A manual inspection indicated that five of these runs are clear outliers. Four of these five runs were measured after a break of several days which seems to have lead to disturbances in chromatography and sample composition. The fifth outlier has a significantly elevated noise level.

Figure 5.12 (a) gives the Mahalanobis distances for this data set. Runs that were classified as outliers by manual inspection are colored in red, normal runs in green. The numbering of data indicates the time order of runs. As we can see, all known outlier maps are recovered. Additionally, some normal runs are classified as mild outliers. Due to the complex composition of the samples, it is difficult to judge whether these runs comprise true outliers that were not discovered during the manual inspection or not. In a real-world study without enough time to perform a manual validation, one would discard the strong (and true) outlier maps. Depending on time and lab resources available, we would recommend to treat the mild outliers with caution or even to repeat these experiments.

Figure 5.12 (b) shows a biplot of the robust PCA analysis. This plot shows that the data (scores) projected on the first two principal components and the variables (loadings) plotted as arrows. This plot exhibits several interesting features. First, we see that the LC-MS maps fall into two distinct clusters. In fact, the maps in the cluster in the upper left half are all maps from the first half of the study whereas the second cluster consists only of maps from the second half. All outliers identified in the previous plot are colored in red and numbered. We can see that outlier 51 stands out because of skewness, kurtosis and maximum intensity value of its TIC. The remaining outliers differ from the rest of the maps mainly because of their elevated intensity values and higher noise content.

Note that the biplot exhibits one potential weakness of the Mahalanobis distance if used for outlier detection: the data is expected to form a single cluster and the method gives less reliable results of this is not the case. This is also the reason why several well reproducible runs were classified as mild outliers.



(a) Mahalanobis distances

(b) Biplot

Figure 5.12: *(a) Log-scaled Mahalanobis distances for the renal patient samples. The black line gives a cutoff for a significance level of 5%. (b) Projection of the LC-MS maps on their first two principal components. Manually identified outliers are colored in red.*

5.5 Discussion and Summary

Quality assessment and control are common place in fields where many items are produced at a rapid pace and where quality is crucial: be it tools in factories or data in high-throughput biological experiments. The application of statistical quality assessment to quantitative mass spectrometry data is still an underexplored field. We expect that, with the growth of this field, this is going to change as much as it has changed for gene expression studies.

It is clearly desirable to have algorithms for automatic quality assessment of samples since in high-throughput studies, manual validation would be a significant bottleneck. That is why we developed a method for the detection of outlier experiments in large-scale mass spectrometry studies. Our method uses different variables to describe an LC-MS experiment. These descriptors capture different aspects of intensity, m/z and noise level distribution of all spectra in an LC-MS map. Using these descriptors we can represent an LC-MS map as a vector of its descriptors. We apply a robust principal component analysis and the Mahalanobis distance to find outlier runs within large numbers of LC-MS maps where each map is represented by its descriptors. The Mahalanobis distance is a variant of the Euclidean distance and weights the distance in each dimension by the variation of this dimension. This weighting is represented by the inverted covariance matrix.

We can use the Mahalanobis distance to compute the distance between two objects but also the distance between one object and a set of objects where this set is represented by its empirical mean and variance. The robust principal component analysis (robust PCA) helps to deal with cases where we have less maps than quality descriptors. In this case, the covariance matrix is singular and cannot be inverted as required for the computation of the Mahalanobis distance. We apply a robust version of principal component analysis since PCA is prone to yield inaccurate results in the presence of outliers.

We demonstrated that our approach works well with large data sets and can accurately detect poor LC-MS runs. This is of special importance in high-throughput experiments, where many LC-MS maps are generated and the time lacks to perform a manual quality assessment. We evaluated our approach on simulated LC-MS runs and two real data sets consisting of around 50 replicates each. In all cases, we were able to detect outlier data sets, outliers that were be confirmed by manual validation. Furthermore, we applied our method to a set of 9 high-quality LC-MS maps with BSA peptides spiked-in at different concentrations. Our method does not classify any of these maps as outlier. This proves that our method is highly robust and cannot only be applied to assess the quality of sets of replicates but also in differential expression studies.

When dealing with outliers, we have two choices: to either remove them or to repeat the corresponding LC-MS run. Clearly, this depends on time and lab resources available. In each case, outlier detection and removal as early as possible during the data analysis will make the results more reliable and save a lot of time.

There are, of course, aspects of this approach that could be improved. For instance, LC-MS experiments that form two or more cluster in the principal component space pose a problem. This is a deficiency inherent to the Mahalanobis distance and might or might not be common for LC-MS data. Nevertheless, we will explore avenues to remedy this problem.

Chapter 6

Summary and Outlook

Synopsis: *We summarize the results of this thesis and give an overview of possible extensions.*

6.1 Novel Contributions of this Thesis

This thesis contains three novel contributions: first, we presented SweepWavelet, an algorithm for the computational quantification of peptides from LC-MS data (Schulz-Trieglaff *et al.*, 2007). We introduced this algorithm in Chapter 3. It includes three novel aspects as compared to previous approaches: a novel wavelet function to filter mass spectra for peptide signals, a method inspired by the sweep-line algorithm to efficiently summarize isotope patterns in adjacent spectra and finally, a highly flexible two-dimensional model to filter the extracted signals for false positives. In contrast to competing approaches, our model incorporates parameters for mass resolution and the elution behavior of the peptide and can be applied to LC-MS data from various instruments. We showed that we can accurately detect and quantify peptides over a large range of mass resolutions and in complex samples.

Second, we performed benchmark studies and defined quality metrics for the comparison of quantification algorithms (Schulz-Trieglaff *et al.*, 2008a,b). We presented the results in Chapter 4. To our knowledge, it is the first time that such a comparison has been performed. Benchmarking LC-MS data analysis algorithms is inherently difficult since the ground truth for a real LC-MS map is usually not known. In an ideal world, this ground truth would consist of a list of all the peptide signals together with their positions in m/z, rt and their relative abundances. Consequently, we used LC-MS data sets which were either manually annotated or generated from a sample of known composition. A manual annotation is of course tedious and only feasible for simple data sets. Furthermore, even for a sample of known composition, it is usually not clear which peptides will actually appear in the LC-MS map. To circumvent these problems, we developed a simulation engine for LC-MS data, called LC-MSsim. Using this software, we can simulate LC-MS maps of different complexities. The user can choose a set of proteins, noise level, chromatographic conditions and mass resolution. LC-MSsim performs a computational digestion of the proteins and determines for each peptide whether it will appear in the LC-MS map. We implemented models for retention time prediction, chromatographic elution peaks and shot noise. The result of a simulation is an LC-MS map, written to an open XML format and, most important for benchmarks, a list of the positions in m/z and rt of each peptide, together with the simulated charge state and the amino acid sequence. Using these lists, we can match the results of a feature detection algorithm to the simulated peptide lists and determine accurate error rates. LC-MSsim is an open source software, freely available to the community, and part of the OpenMS software library. We are not aware of any other software that can simulate a full LC-MS setup.

Our third contribution (Schulz-Trieglaff *et al.*, 2008c) is a statistical model for the quality assessment of large-scale LC-MS experiments in Chapter 5. While quality assessment and control are common in other fields, such as sequence analysis and gene expression measurements, it has not yet been widely applied to LC-MS experiments. Our approach relies on a set of quality descriptors to describe an LC-MS map. These descriptors capture various aspects of the quality of an LC-MS map such as noise level or the shape of the total ion chromatogram. With these descriptors, we can describe an LC-MS map using a vector where each entry corresponds to a descriptor. We use an outlier detection method based on the Mahalanobis distance. If we applied this algorithm directly, we would face two problems: first, the method is not robust and might lead to incorrect results if the data contains many outlying observations. Second, the covariance matrix needed to compute the Mahalanobis distance, is singular if the number of experiments is smaller than than the number of quality descriptors. To solve these problems, we apply an algorithm based on projection pursuit and compute a robust version of principal component analysis. The result is a smaller list of quality descriptors for each LC-MS map, where each new descriptor is now a linear combination of the previous ones. We evaluate our method on simulated and real data. We show that it is superior to standard outlier detection methods which do not rely on robust estimators, and that we can accurately find outlier in large-scale experimental studies.

6.2 Future Work

Each of the three novel contributions of this thesis have reached a mature stage of development and have been used in real world projects. But there is always room for improvement. To give an example, since the first publication (Schulz-Trieglaff *et al.*, 2007) of our quantification algorithm SweepWavelet, several new algorithms have been developed. Nevertheless, there is currently no algorithm that supports a quantification using labelling technologies, such as ICAT or SILAC, in a straightforward manner. Of course, all classical algorithms support a quantification using mass-tag labels by performing a search for pairs of peptide features with a given mass distance. But no algorithm incorporates this step directly into the feature discovery process.

Remember that the application of mass-tag labels for quantification results in pairs of peptide features with a defined mass distance in the same LC-MS map. These pairs are from different samples such as diseased versus healthy tissue. Only one sample will be treated with a heavier isotope or some other chemical modification, leading to the aforementioned mass differences. In the same manner as the isotope wavelet is tailored for the detection of isotope pattern, it should be possible to design a method that explicitly spectra spectra for pairs of isotope signals with a given mass distance and suppresses all other signals. This would greatly reduce the number of false-positives in a labelled quantification experiment. The method does not need to use wavelets, a simpler matched filter might suffice. A first step into this direction has already been taken (McIlwain *et al.*, 2007). Nevertheless, their method relies on manually annotated peak lists to achieve acceptable error rates and is thus of less practical use.

One might use even more additional information during the peptide feature detection process. To give an example, one could combine MS/MS sequence and feature information. If a high-quality feature is detected, but the associated MS/MS spectrum could not be identified, one might perform a more time-consuming identification such as a database search incorporating modifications or de-novo sequencing. On the other hand, if a MS/MS spectrum could be matched with a peptide sequence but no MS feature was found at this position, one might re-iterate the feature detection process. This idea could be extended even further: on high-resolution spectra, we could use information from MS peptide features to alleviate the identification of overlapping or modified peptides.

Finally, applications of our quantification algorithm to real world data showed that parameter tuning of the algorithm to LC-MS data from a particular instrument can be very time consuming. Parameters that need to be adopted are width of the isotope peaks, shape of the elution profile, etc. Manual fine-tuning of these parameters is laborious. It might be worthwhile to explore a semi-manual parameter learning strategy: the user could manually annotate a small set of peptide signals in a representative LC-MS map. The algorithm could learn a set of suitable parameters from this feature set and perform feature detection using these parameters on the whole data set.

Regarding the benchmarking chapter of this thesis, our software LC-MSsim is available under an open source license and we hope that it will be used and improved in the future.

An obvious extension would be the simulation of MS/MS spectra. While the enumeration of the theoretically possible MS/MS fragments of a peptide is not difficult, the computational prediction of the fragments that will actually occur in the spectrum and their intensities is far more difficult. It might be useful to add pre-defined parameter sets for certain instruments or LC-MS setups. Furthermore, we believe that there is a pressing need in the mass spectrometry community for a collection of quantitative LC-MS benchmark data. This would be extremely helpful for algorithm developers but also for users who want to find the best algorithm for their respective LC-MS setup. These benchmark data sets might comprise real and simulated data. We believe that our work is a step into the right direction, but more should follow.

Finally, we believe that with the ever increasing number of proteomics experiments and labs, automatic quality assessment methods will become more common. This was also the case for microarray data (Brettschneider *et al.*, 2007) and DNA sequencing (Ewing *et al.*, 1998). We believe that a statistical quality control should be part of every proteomics experiment. In the ideal case, wet lab scientists would even consult a statistician before they perform their experiments, but this is not always done.

The outlier detection method we developed is designed for raw LC-MS maps. Ideally, we would have a quality assessment on every level of the data analysis workflow: raw data, peptide features, alignment, etc. Quality metrics for peptide feature lists exist (Piening *et al.*, 2006) and quality assessment of aligned and de-charged feature maps might be performed in a similar way as for microarrays, but this remains to be tested. There is currently no consistent framework that combines all these methods.

Appendix

6.3 Availability and Implementation

All software tools presented in this thesis were developed using OpenMS, software framework for computational shotgun proteomics written in C++.

6.3.1 Quantification Algorithm

The presented algorithm for peptide quantification, SweepWavelet, is part of OpenMS release version 1.0. In this work, we used a slightly modified version which is stored in a separate branch. The algorithm is also implemented as a stand-alone component of The OpenMS Proteomics Pipeline (TOPP) (Kohlbacher *et al.*, 2007), which is a spin-off project of OpenMS. The source code can be retrieved using:

```
svn co https://open-ms.svn.sourceforge.net/svnroot/open-ms/branches/FF10
```

The installation instructions are the same as for the OpenMS release version 1.1 and can be obtained from http://www.openms.de.

In contrast to the algorithm published in (Schulz-Trieglaff *et al.*, 2007), we do not compute the full two-dimensional model (e.g. m/z and retention time combined). In the current version of our algorithm, we compute the goodness of fit between m/z and retention time separately by first projecting on m/z and then on retention time. This results in a significant speed-up as compared to the full two-dimensional model. The final correlation or goodness-of-fit is then the average of the rank correlation in m/z and rt.

6.3.2 LC-MS Data Simulator

The simulator was developed using the OpenMS library, but at the time this thesis is written, it is not part of the official release version. The source code can be obtained from:

```
svn co https://lcms-sim.svn.sourceforge.net/svnroot/lcms-sim lcms-sim
```

6.3.3 Quality Assessment Software

The software is written in C++ and R. The source code can be obtained from the author on request.

6.4 Parameter Settings for Feature Detection Benchmarks

This text was published in the supplementary material of Schulz-Trieglaff *et al.* (2008b). Several of the quantification tools that we tested depend a lot on parameter tuning. Nevertheless, most parameters are not well documented. On the other hand, some tools work out-of-the-box and do not offer any parameters that could be adopted. We chose the parameter settings for each tool as described below. Three tools (msInspect, Decon2LS and MZmine) still produced large numbers of false positives, as stated below. We therefore removed for each of these three tools all peptide features with an intensity below the first quantile. The first quantile is the value such that 25% of all feature intensities are below it.

6.4.1 msInspect

This tools does not have any parameters besides m/z and rt ranges searched. Each isotope pattern is scored using the Kullback-Leibler (KL) distance between an averagine model and the true peak intensities. A KL distance of zero indicates a perfect match between averagine model and signal and thus a high-quality peptide feature. To find a suitable cutoff for this score, we executed msInspect on a complex LC-MS run, a digest of Halobacterium NRC-1 proteins recorded on a API QSTAR Pulsar I instrument (downloaded from the PeptideAtlas database, ref: Pae000245, data set # 25, http://www.peptideatlas.org/). We manually annotated this data set and choose 20 intense and well-resolved peptide signals to determine a suitable cutoff. We choose a cutoff for the KL distance of 0.8 such that 80% of the annotated features were detected. Applying this filtering threshold to the whole data set, the number of detected features was reduced from 3366 to 2608 features.

6.4.2 Decon2LS

This tool offers plenty of parameters, some are documented and some are not. According to the authors (personal communication), the fit intensity and the fit score threshold should have significant influence on the result. However, we made the experience that the fit intensity threshold has not much influence on the result. Consequently, we optimized the fit score threshold (a distance measure between 0 and 1, where 0 is best) in the same way as the KL distance described above. We choose a cutoff for the fit score of 0.2 which resulted in 85% annotated features recovered. The overall number of features was reduced from 19603 to 14592.

6.4.3 MZmine

This software offers different peak detection strategies. We used the *recursive threshold peak detection* algorithm. Furthermore, the software has several parameters that influence the feature detection process. This process consists of two parts: a peak detection and a de-isotoping step. Peak detection is influenced by parameters such as bin sizes, min intensity etc. The de-isotoping step groups peaks into isotope pattern, estimates a charge and removes incomplete isotope pattern and single peaks. We contacted the developers of MZmine and asked them which parameters would have significant influence. MZmine does not compute a score or s/n threshold that could be used as single filter criterion. According to the recommendations of the MZmine developers, we adopted the *Chromatographic threshold level* and the *m/z bin size* as well as the noise threshold and minimum peak height which are both given in absolute intensity units. For the high-resolution data (FWHM 0.02), we choose a set of parameters that recovered 100% of the annotated features from the QSTAR mass spectrometer. With decreasing mass resolution, we relaxed the bin width and the chromatographic threshold level but without satisfying results. In each case, we estimated the noise level as the median intensity on a small m/z interval in empty region (i.e. without peptide signals) of the LC-MS map. We set the minimum peak intensity to the same value.

6.4.4 SpecArray

No parameter changes possible, no manual tuning.

6.4.5 SweepWavelet / Superhirn

Several parameter offered, moderate optimization to achieve trade-off between false positives and false negatives.

Glossary

Base Peak

The most intense ion detected in a mass spectrum, used to construct the total ion chromatogram (TIC).

Capillary Electrophoresis-Mass Spectrometry (CE-MS)

The linking of mass spectrometry to a capillary. This method separates charged molecules based on their mass-to-charge ratio and frictional forces. The effluent from the capillary is introduced directly into the mass spectrometer source.

Electrospray Ionization (ESI)

An ionization method frequently appled in conjunction with liquid chromatography. The sample compounds are ionized from an aqueous solution by pushing it through a capillary tube at a high potential. ESI results in several charge variants per compounds. Depending on the size of the molecule, very high charge states ($geq20$) can be observed.

False Discovery Rate (FDR)

A method to deal with the problem of multiple hypothesis testing. The FDR is the expected proportion of incorrectly rejected null hypotheses (i.e. false positives). It is a relatively liberal metric to control the number of false positives and can be derived from the observed distribution of p-values.

Full Width at Half Maximum (FWHM)

The full width at half maximum (FWHM) is a unitless measure of the spread of a function or signal. In the case of a mass spectral peak, it is given by the difference between the two m/z values at which the intensity is equal to half of its maximum.

Isotopes

The isotopes of an elements have the same number of protons but differ in the number of neutrons and thus in mass. Most elements are composed of several isotopes or isotope variants. If separated by a mass spectrometer, the isotopes will display a cluster of ions, the iosotopic pattern.

Isotopic Pattern

A group of peaks. More precisely, a group of peaks all caused by the same peptide ion and exhibiting a distinctive intensity pattern due to the atomic composition of the peptide.

Liquid Chromatography-Mass Spectrometry (LC-MS)

The linking of the effluent from a liquid chromatographic system to a mass spectrometer. This setup is used for high-throughput proteomics experiments.

LC-MS Map
The set of spectra obtained from an LC-MS experiment. Each spectrum represents a subset of the sample compounds as they elute from the mass spectrometer.

Mass Defect
This term describes the observation that the measured mass of an atom is less than the sum of the masses of its protons, neutrons, and electrons. It is due to the binding energy of the nucleus.

Mass Deviance
The minimum distance between the mass of a feature and the closest peptide mass obtained from a theoretical protein digest.

Mass Resolution
(Informally) defined as the Full-Width-At-Half-Maximum of a peak divided by its mass.

Mass Spectrum
A plot of m/z (abscissa) versus the intensity (ordinates). The spectrum is produced by scanning the analyzer to transmit ions (or release them from a trap) for a predefined range of m/z values over a fixed period of time. Thomson is the unit for mass-to-charge ratio.

Matrix-Assisted Laser Desorption/Ionization(MALDI)
An alternative ionization method to ionize molecules out of a dry, crystalline matrix via laser pulses. The liquid sample is mixed with so-called matrix molecules. The solution is spotted on the MALDI plate and dried such that the matrix molecules crystallize. The sample analyte is captured between the matrix molecules and ionized by a laser beam. The matrix protects the analyte and facilitates the ionization.

Monoisotopic Ion Mass
The mass of the ion containing the most abundant isotopes, calculated with exact atomic weights.

Parent Ion
A signal in an MS spectrum which triggers the acquisition of an MS/MS spectrum. Also called precursor ion.

Peak
A local and pronounced (i.e. significantly higher than background noise) maximum in a mass spectrum.

Peptide Feature
All data points in an LC-MS map caused by a charge variant of a peptide.

Precursor Ion
See parent ion.

Proteomics
The systematic study of the abundances, sequences, chemical modifications, structure and localization of all proteins contained in a living system, such as a human cell or a whole organism. This term was coined to make an analogy with genomics.

Tandem Mass Spectrometry (MS/MS)

The aim of this experiment is to generate fragments of selected (precursor) ions. To achieve this, two mass analyzers are coupled and separated by a region in which the ions are fragmented by transfer of energy (usually by collision with other molecules). The fragments are recorded in a mass spectrum and can be used to infer the identity of the molecule (or sequence in the case of peptides).

Time-of-Flight (TOF)

A method to measure the mass-to-charge ratio of ions. It can be used with ESI (ESI-TOF) or MALDI (MALDI-TOF). The ions are accelerated by an electric field of known strength and the time that they need to reach a detector through a vacuum is measured. Since their velocity depends on the mass-to-charge ratio, it can be readily deduced from the flight time.

Total Ion Chromatogram

The total ion chromatogram (TIC) is the projection of an LC-MS map on its retention time axis. In general, for each spectrum its sum of point intensities is plotted but it can also be construted using the base peak, i.e. the most intense peak in each spetrum.

Index

Curriculum Vitae

Der Lebenslauf ist in der Online-Version aus Griünden des Datenschutzes nicht enthalten.

Der Lebenslauf ist in der Online-Version aus Griünden des Datenschutzes nicht enthalten.

Bibliography

Aldroubi, A. and Unser, M. (1996). *Wavelets in Medicine and Biology*. CRC Press, Inc.

Alsberg, B. K., Woodward, A. M., and Kell, D. B. (1997). An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemometrics and Intelligent Laboratory Systems*, **37**(2), 215–239.

Anderle, M., Roy, S., Lin, H., Becker, C., and Joho, K. (2004). Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics*, **20**, 3575–3582.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition.

Baggerly, K. A., Morris, J. S., Wang, J., Gold, D., Xiao, L. C., and Coombes, K. R. (2003). A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Protemics*, **3**, 1667–1672.

Bandeira, N., Tsur, D., Frank, A., and Pevzner, P. A. (2007a). Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences*, **104**(15), 6140–6145.

Bandeira, N., Clauser, K. R., and Pevzner, P. A. (2007b). Shotgun Protein Sequencing: Assembly of Peptide Tandem Mass Spectra from Mixtures of Modified Proteins. *Mol Cell Proteomics*, **6**(7), 1123–1134.

Bellew, M., Coram, M., Fitzgibbon, M., Igra, M., Randolph, T., Wang, P., May, D., Eng, J., Fang, R., Lin, C.-W., Chen, J., Goodlett, D., Whiteaker, J., Paulovich, A., and McIntosh, M. (2006). A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, **22**(15), 1902–1909.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Bern, M., Goldberg, D., McDonald, W. H., and Yates, John R., I. (2004). Automatic Quality Assessment of Peptide Tandem Mass Spectra. *Bioinformatics*, **20**, i49–54.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60.

Boone, B., Mitchum, R. K., and Schsppele, S. E. (1970). Computer analysis of low resolution mass spectra correction for natural abundance of 13c, 2h, 15n, 17o and 18o. *International Journal or Mass Spectrometry and Ion Physics*, **5**(1-2), 21–27.

Bose, P., Maheshwari, A., and Morin, P. (2002). Fast approximations for sums of distances, clustering and the fermatweber problem. *Computational Geometry*, **24**(3), 135–146.

Brettschneider, J., Collin, F., Bolstad, B. M., and Speed, T. P. (2007). Quality assessment for short oligonucleotide microarray data. *Technometrics*.

Brown, C. S., Goodwin, P. C., and Sorger, P. K. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences*, **98**(16), 8944–8949.

Brownawell, M. L. and San Filippo, Joseph, J. (1982). A program for the synthesis of mass spectral isotopic abundances. *Journal of Chemical Education*, **59**(8), 663–5.

Carrick, A. and Glockling, F. (1967). Mass and abundance data for polyisotopic elements. *Journal of the Chemical Society A*, pages 40–42.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chankvetadze, B. (1997). *Capillary Electrophoresis in Chiral Analysis*. John Wiley and Sons.

Chen, L. and Yap, Y. L. (2008). Automated charge state determination of complex isotope-resolved mass spectra by peak-target fourier transform. *Journal of the American Society for Mass Spectrometry*, **19**(1), 46–54.

Chen, L., Sze, S., and Yang, H. (2006). Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. *Analytical Chemistry*, **78**(14), 5006–5018.

Chernushevich, I. V., Loboda, A. V., and Thomson, B. A. (2001). An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of Mass Spectrometry*, **36**(8), 849–865.

Choo, K. and Tham, W. (2007). Tandem mass spectrometry data quality assessment by self-convolution. *BMC Bioinformatics*, **8**(1), 352.

Cohen Freue, G. V., Hollander, Z., Shen, E., Zamar, R. H., Balshaw, R., Scherer, A., McManus, B., Keown, P., McMaster, W. R., and Ng, R. T. (2007). MDQC: a new quality assessment method for microarrays based on quality control reports. *Bioinformatics*, **23**(23), 3162–3169.

Comisarow, M. B. and Marshall, A. G. (1974). Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters*, **25**, 282–283.

Coombes, K. R., Fritsche, Herbert A., J., Clarke, C., Chen, J.-n., Baggerly, K. A., Morris, J. S., Xiao, L.-c., Hung, M.-C., and Kuerer, H. M. (2003). Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization. *Clin Chem*, **49**(10), 1615–1623.

Coombes, K. R., Koomen, J., Baggerly, K. A., Morris, J. S., and Kobayashi, R. (2005). Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics*, **1**(1).

Crick, F. (1970). The central dogma of molecular biology. *Nature*, **227**, 561–563.

Croux, C. and Ruiz-Gazen, A. (1996). A fast algorithm for robust principal components based on projection pursuit. In A. Prat, editor, *COMPSTAT: Proceedings in Computational Statistics*, page 211216. Physica-Verlag.

Daszykowski, M., Kaczmarek, K., Heyden, Y. V., and Walczaka, B. (2007). Robust statistics in data analysis a review basic concepts. *Chemometrics and Intelligent Laboratory Systems*, **85**, 203–219.

de Hoffmann, E. and Stroobant, V. (2004). *Mass Spectrometry: Principles and Applications*. Wiley & Sons.

Decramer, S., Gonzalez de Peredo, A., Breuil, B., Mischak, H., Monsarrat, B., Bascands, J.-L., and Schanstra, J. P. (2008). Urine in clinical proteomics. *Mol Cell Proteomics*, pages R800001–MCP200.

Di Marco, V. B. and Bombi, G. G. (2001). Mathematical functions for the representation of chromatographic peaks. *Journal of Chromatography A*, **931**, 1–30.

Du, P. and Angeletti, R. (2006). Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Analytical Chemistry*, **78**(10), 3385–3392.

Du, P., Sudha, R., Prystowsky, M. B., and Angeletti, R. H. (2007). Data reduction of isotope-resolved lc-ms spectra. *Bioinformatics*, **23**(11), 1394–1400.

Du, P., Stolovitzky, G., Horvatovich, P., Bischoff, R., Lim, J., and Suits, F. (2008). A Noise Model for Mass Spectrometry Based Proteomics. *Bioinformatics*, pages 1070–1077.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred I. Accuracy Assessment. *Genome Res.*, **8**(3), 175–185.

Fenn, J., Mann, M., Meng, C., Wong, S., and Whitehouse, C. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science*, **246**(4926), 64–71.

Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K., and Eidhammer, I. (2006). Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, **6**(7), 2086–2094.

Fraser, A., Hengartner, N., Vixie, K., and Wohlberg, B. (2003). Incorporating invariants in mahalanobis distance based classifiers: application to face recognition. *Proceedings of the International Joint Conference on Neural Networks*, **4**, 3118– 3123.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, **23**, 881890.

Gay, S., Binz, P.-A., Hochstrasser, D. F., and Appel, R. D. (2002). Peptide mass fingerprinting peak intensity prediction: Extracting knowledge from spectra. *Proteomics*, **2**(10), 1374–1391.

Ge, G. and Wong, G. W. (2008). Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, **9**(1), 275.

Gilbert, W. (1986). Origin of life: The rna world. *Nature*, **319**(6055), 618–618.

Graps, A. (1995). Introduction to wavelets. *IEEE Computational Science and Engineering*, **2**(2), 50–61.

Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R. (2002). Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in Saccharomyces cerevisiae. *Mol Cell Proteomics*, **1**(4), 323–333.

Gröpl, C., Lange, E., Reinert, K., Kohlbacher, O., Sturm, M., Huber, C. G., Mayr, B., and Klein, C. (2005). Algorithms for the automated absolute quantification of diagnostic markersin complex proteomics samples. In M. Berthold, editor, *Procceedings of CompLife 2005*, Lecture Notes in Bioinformatics, pages 151–163. Springer, Heidelberg.

Grunert, M. (2008). *Algorithmic improvements and automated parameter learning for feature finding in LC-MS and CE-MS data*. Master's thesis, Deparment of Computer Science and Mathematics, Free University of Berlin.

Grushka, E. (1972). Characterization of exponentially modified gaussian peaks in chromatography. *Anal. Chem.*, **44**(11), 1733–1738. First peak on application of EMG for elution profiles.

Henry, K. D. and McLafferty, F. W. (1990). Electrospray ionization with fourier-transform mass spectrometry. charge state assignment from resolved isotopic peaks. *Organic mass spectrometry*, **25**(9), 490–492.

Hill, E. G., Schwacke, J. H., Comte-Walters, S., Slate, E. H., Oberg, A. L., Eckel-Passow, J. E., Therneau, T. M., and Schey, K. L. (2008). A statistical model for itraq data analysis. *Journal of Proteome Research*.

Hoopmann, M., Finney, G., and MacCoss, M. (2007). High-speed data reduction, feature detection, and ms/ms spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Analytical Chemistry*, **79**(15), 5620–5632.

Horn, D. M., Zubarev, R. A., and McLafferty, F. W. (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, **11**(4), 320–332.

Hössjer, O. and Croux, C. (1995). Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *Journal of Nonparametric Statistics*, **4**(3), 293–308.

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Cooks, R. G. (2005). The orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, **40**(4), 430–443.

Hussong, R., Tholey, A., and Hildebrandt, A. (2007). Efficient analysis of mass spectrometry data using the isotope wavelet. In *COMPLIFE : The Third International Symposium on Computational Life Science. American Institute of Physics (AIP) Proceedings Volume 940*, pages 139–149.

Iavarone, A. T., Jurchen, J. C., and Williams, E. R. (2000). Effects of solvent on the maximum charge state and charge state distribution of protein ions produced by electrospray ionization. *Journal of the American Society for Mass Spectrometry*, **11**(11), 976–985.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucl. Acids Res.*, **31**(4), e15–.

Jaffe, J. D., Berg, H. C., and Church, G. M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. *PROTEOMICS*, **4**(1), 59–77.

Kagen, L., Scheidt, S., Roberts, L., Porter, A., and Paul, H. (1975). Myoglobinemia following acute myocardial infarction. *The American Journal of Medicine*, **58**(2), 177–182.

Kasicka, V. (2001). Recent advances in capillary electrophoresis of peptides. *ELECTROPHORESIS*, **22**(19), 4139–4162.

Katajamaa, M. and Orešič, M. (2005). Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, **6**, 179.

Katajamaa, M., Miettinen, J., and Orešič, M. (2006). MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, **22**(5), 634–636.

Kaur, P. and O'Connor, P. (2006). Comparison of charge state determination methods for high resolution mass spectra. In *2006 IEEE International Conference on Granular Computing*.

Kaur, P. and OConnor, P. B. (2006). Algorithms for automatic interpretation of high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, **17**, 459–68.

Klammer, A., Yi, X., MacCoss, M., and Noble, W. (2007). Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Analytical Chemistry*, **79**(16), 6111–6118.

Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., and Sturm, M. (2007). TOPP–The OpenMS Proteomics Pipeline. *Bioinformatics*, **23**(2), e191–197.

Konermann, L. (2007). A minimalist model for exploring conformational effects on the electrospray charge state distribution of proteins. *Journal of Physical Chemistry B*, **111**(23), 6534–6543.

Koornwinder, T. H. (1998). *Wavelets: An Elementary Treatment of Theory and Applications*. World Scientific.

Krokhin, O. V. (2006). Sequence-specific retention calculator. algorithm for peptide retention prediction in ion-pair rp-hplc: application to 300- and 100-a pore size c18 sorbents. *Anal Chem*, **78**(22), 7785–7795.

Kubinyi, H. (1991). Calculation of Isotope Distributions in Mass Spectrometry. A Trivial Solution for a Non-Trivial Problem. *Anal. Chim. Acta*, **247**, 107 – 109.

Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., and Aebersold, R. (2007). Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, **7**(5), 655–667.

Lange, E., Gröpl, C., Schulz-Trieglaff, O., Leinenbach, A., Huber, C., and Reinert, K. (2007). A geometric approach for the alignment of Liquid Chromatography Mass Spectrometry data. *Bioinformatics*, **23**(13), i273–281.

Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M. W., Pain, A., Sauerwein, R. W., Eling, W. M. C., Hall, N., Waters, A. P., Stunnenberg, H. G., and Mann, M. (2002). Analysis of the plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature*, **419**(6906), 537–542.

Le Texier, V., Riethoven, J.-J., Kumanduri, V., Gopalakrishnan, C., Lopez, F., Gautheret, D., and Thanaraj, T. A. (2006). Alttrans: Transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics*, **7**(1), 169.

Lee, C. and Roy, M. (2004). Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biology*, **5**(7), 231.

Leptos, K. C., Sarracino, D. A., Jaffe, J. D., Krastins, B., and Church, G. M. (2006). MapQuant: Open-Source software for large-scale protein quantification. *Proteomics*, **6**(6), 1770–1782.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quart. Appl. Math.*, **2**, 164–168.

Li, G. and Chen, Z. (1985). Robust projection pursuit estimator for dispersion matrices and principal components. *Journal of the American Statistical Association*, **80**(391), 759–766.

Li, J. (2002). Comparison of the capability of peak functions in describing real chromatographic peaks. *Journal of Chromatography A*, **952**(1-2), 63–70.

Li, X.-j., Yi, E. C., Kemp, C. J., Zhang, H., and Aebersold, R. (2005). A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry. *Mol Cell Proteomics*, **4**(9), 1328–1340.

Listgarten, J. and Emili, A. (2005). Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics*, **4**(4), 419–434.

Listgarten, J., Neal, R. M., Roweis, S. T., Wong, P., and Emili, A. (2007). Difference detection in lc-ms data for protein biomarker discovery. *Bioinformatics*, **23**, e198–e204.

Liu, H., Sadygov, R., and Yates, J. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry*, **76**(14), 4193–4201.

Lopuhaa, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, **19**(1), 229–248.

MacCoss, M. and Matthews, D. E. (2005). Quantitative MS for proteomics: Teaching a new dog old tricks. *Anal. Chem.*, **77**(15), 294A–302A.

Machtejevas, E., Andrecht, S., Lubda, D., and Unger, K. K. (2007). Monolithic silica columns of various format in automated sample clean-up/multidimensional liquid chromatography/mass spectrometry for peptidomics. *Journal of Chromatography A*, **1144**(1), 97–101.

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, **12**, 49–55.

Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., and Aebersold, R. (2007). Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotech*, **25**(1), 125–131.

Mann, M. and Aebersold, R. (2003). Mass spectrometry-based proteomics. *Nature 422*, **422**, 198 – 207.

Mann, M., Meng, C. K., and Fenn, J. B. (1989). Interpreting mass spectra of multiply charged ions. *Analytical Chemistry*, **61**(15), 1702–1708.

March, R. E. (2000). Quadrupole ion trap mass spectrometry: a view at the turn of the century. *International Journal of Mass Spectrometry*, **200**, 85–312.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, **11**, 431–441.

Mayr, B. M., Kohlbacher, O., Reinert, K., Sturm, M., Gröpl, C., Lange, E., Klein, C., and Huber, C. (2006). Absolute myoglobin quantitation in serum by combining two-dimensional liquid chromatography-electrospray ionization mass spectrometry and novel data analysis algorithms. *J. Proteome Res.*, **5**, 414–421.

McIlwain, S., Page, D., Huttlin, E. L., and Sussman, M. R. (2007). Using dynamic programming to create isotopic distribution maps from mass spectra. *Bioinformatics*, **23**(13), i328–336.

Meek, J. L. (1980). Prediction of Peptide Retention Times in High-Pressure Liquid Chromatography on the Basis of Amino Acid Composition. *PNAS*, **77**(3), 1632–1636.

Mitchell Wells, J. and McLuckey, S. A. (2005). Collision[hyphen (true graphic)]induced dissociation (cid) of peptides and proteins. In A. L. Burlingame, editor, *Biological Mass Spectrometry*, volume Volume 402, pages 148–185. Academic Press.

Model, F., Konig, T., Piepenbrock, C., and Adorjan, P. (2002). Statistical process control for large scale microarray experiments. *Bioinformatics*, **18**, S155–163.

Moore, R. E., Young, M. K., and Lee, T. D. (2000). Method for screening peptide fragment ion mass spectra prior to database searching. *Journal of the American Society for Mass Spectrometry*, **11**(5), 422–426.

Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.-Y., Vitek, O., Aebersold, R., and Müller, M. (2007). Superhirn - a novel tool for high resolution lc-ms-based peptide/protein profiling. *Proteomics*, **7**(19), 3470–3480.

Na, S. and Paek, E. (2006). Quality assessment of tandem mass spectra based on cumulative intensity normalization. *Journal of Proteome Research*, **5**(12), 3241–3248.

Naish, P. and Hartwell, S. (1988). Exponentially Modified Gaussian functions: A good model for chromatographic peaks in isocratic HPLC? *Chromatographia*, **26**(1), 285–296.

Nelson, T. J., Backlund, Peter S., J., Yergey, A. L., and Alkon, D. L. (2002). Systematic identification of protein-protein interactions by mass spectrometry. *Mol Cell Proteomics*, pages T100006–MCP200.

Nocedal, J. and Wright, S. J. (2006). *Numerical optimization*. Springer.

Oberg, A. L., Mahoney, D. W., Eckel-Passow, J. E., Malone, C. J., Wolfinger, R. D., Hill, E. G., Cooper, L. T., Onuma, O. K., Spiro, C., Therneau, T. M., and Bergen, III, H. R. (2008). Statistical analysis of relative labeled mass spectrometry data from complex samples using anova. *Journal of Proteome Research*, **7**(1), 225–233.

Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., and Ahn, N. G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics*, **4**(10), 1487–1502.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002). Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol Cell Proteomics*, **1**(5), 376–386.

Paul, W. and Steinwedel, H. (1953). Ein neues massenspektrometer ohne magnetfeld. *Zeitschrift fr Naturforschung A*, **8**(7), 448–450.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**, 559572.

Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Petritis, K., Kangas, L. J., Yan, B., Monroe, M. E., Strittmatter, E. F., Qian, W.-J., Adkins, J. N., Moore, R. J., Xu, Y., Lipton, M. S., Camp, D. G., and Smith, R. D. (2006). Improved peptide elution time prediction for reversed-phase liquid chromatography-ms by incorporating peptide sequence information. *Anal Chem*, **78**(14), 5026–5039.

Pfeifer, N., Leinenbach, A., Huber, C. G., and Kohlbacher, O. (2007). Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. *BMC Bioinformatics*, **8**(1), 468.

Piening, B., Wang, P., Bangur, C., Whiteaker, J., Zhang, H., Feng, L.-C., Keane, J., Eng, J., Tang, H., Prakash, A., McIntosh, M., and Paulovich, A. (2006). Quality control metrics for LC-MS feature detection tools demonstrated on saccharomyces cerevisiae proteomic profiles. *Journal of Proteome Research*, **5**(7), 1527–1534.

Prakash, A., Mallick, P., Whiteaker, J., Zhang, H., Paulovich, A., Flory, M., Lee, H., Aebersold, R., and Schwikowski, B. (2006). Signal Maps for Mass Spectrometry-based Comparative Proteomics. *Mol Cell Proteomics*, **5**(3), 423–432.

Prakash, A., Piening, B., Whiteaker, J., Zhang, H., Shaffer, S. A., Martin, D., Hohmann, L., Cooke, K., Olson, J. M., Hansen, S., Flory, M. R., Lee, H., Watts, J., Goodlett, D. R., Aebersold, R., Paulovich, A., and Schwikowski, B. (2007). Assessing bias in experiment design for large-scale mass spectrometry-based quantitative proteomics. *Mol Cell Proteomics*, pages M600470–MCP200.

Prince, J. and Marcotte, E. (2006). Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry*, **78**(17), 6140–6152.

R. Sarpeshkar, T. D. and Mead, C. A. (1993). White noise in MOS transistors and resistors. *IEEE Circuits Devices Mag.*, pages 23–29.

Rappsilber, J. and Mann, M. (2002). Is mass spectrometry ready for proteome-wide protein expression analysis? *Genome Biology*, **3**(8), 2008.1–2008.5.

Reinhold, B. B. and Reinhold, V. N. (1992). Electrospray ionization mass spectrometry: Deconvolution by an entropy-based algorithm. *Journal of the American Society for Mass Spectrometry*, **3**(3), 207–215.

Rockwood, A. and Van Orden, S. (1996). Ultrahigh-speed calculation of isotope distributions. *Analytical Chemistry*, **68**(13), 2027–2030.

Rockwood, A. L., Van Orden, S. L., and Smith, R. D. (1995). Rapid calculation of isotope distributions. *Anal. Chem.*, **67**(15), 2699–2704.

Rodriguez, J., Gupta, N., Smith, R. D., and Pevzner, P. A. (2008). Does trypsin cut before proline? *Journal of Proteome Research*, **7**(1), 300–305.

Sanders, W., Bridges, S., McCarthy, F., Nanduri, B., and Burgess, S. (2007). Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics*, **8**(Suppl 7), S23.

Sauve, A. and Speed, T. (2004). Normalization, baseline correction and alignment of high-throughput mass spectrometry data. In *Proceedings Gensips*.

Schley, C., Swart, R., and Huber, C. (2006). Capillary scale monolithic trap column for desalting and preconcentration of peptides and proteins in one- and two-dimensional separations. *J Chromatogr A*, **1136**(2), 210–220.

Schnier, P. D., Gross, D. S., and Williams, E. R. (1995). On the maximum charge state and proton transfer reactivity of peptide and protein ions formed by electrospray ionization. *Journal of the American Society for Mass Spectrometry*, **6**(11), 1086–1097.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, **12**(5), 1207–1245.

Schulz-Trieglaff, O., Hussong, R., Gröpl, C., Hildebrandt, A., and Reinert, K. (2007). A fast and accurate Algorithm for the Quantification of peptides from LC-MS data. In T. P. Speed and H. Huang, editors, *Research in Computational Molecular Biology, 11th Annual International Conference, RECOMB 2007, Oakland, CA, USA, April 21-25, 2007*, volume 4453 of *Lecture Notes in Computer Science*, pages 473–487. Springer.

Schulz-Trieglaff, O., Hussong, R., Gröpl, C., Leinenbach, A., Hildebrandt, A., Huber, C., and Reinert, K. (2008a). Computational Quantification of Peptides from LC-MS data. *Journal of Computational Biology*.

Schulz-Trieglaff, O., Pfeifer, N., Groepl, C., Kohlbacher, O., and Reinert, K. (2008b). LC-MSsim : a simulation software for Mas Spectrometry-Liquid Chromatography Experiments. *BMC Bioinformatics*.

Schulz-Trieglaff, O., Machtejevas, E., Reinert, K., Schlueter, H., Thiemann, J., and Unger, K. (2008c). Statistical Quality Assessment and Outlier Detection for Liquid Chromatography-Mass Spectrometry Experiments. *BioDataMining*.

Schweppe, R., Haydon, C., Lewis, T., Resing, K., and Ahn, N. (2003). The characterization of protein post-translational modifications by mass spectrometry. *Acc. Chem. Res.*, **36**(6), 453–461.

Scigelova, M. and Makarov, A. (2006). Orbitrap mass analyzer - overview and applications in proteomics. *Proteomics*, **6**(S2), 16–21.

Senko, M., Beu, S., and McLafferty, F. (1995a). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, **6**, 229–233.

Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995b). Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *Journal of the American Society for Mass Spectrometry*, **6**(1), 52–56.

Shiio, Y. and Aebersold, R. (2006). Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry. *Nat. Protocols*, **1**(1), 139–145.

Shin, H., Koomen, J., Baggerly, K., and Markey, M. (2004). Towards a noise model of maldi tof spectra. In *American Association for Cancer Research (AACR) advances in proteomics in cancer research, Key Biscayne, FL.*

Silva, J., Denny, R., Dorschel, C., Gorenstein, M., Kass, I., Li, G.-Z., McKenna, T., Nold, M., Richardson, K., Young, P., and Geromanos, S. (2005). Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.*, **77**(7), 2187–2200.

Smith, L. I. (2002). A tutorial on principal components analysis. Technical report, Department of Computer Science, University of Otago.

Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. (2002). An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, **2**(5), 513–523.

Snider, R. K. (2007). Efficient calculation of exact mass isotopic distributions. *Journal of the American Society for Mass Spectrometry*, **18**(8), 1511–1515.

Stead, D. A., Paton, N. W., Missier, P., Embury, S. M., Hedeler, C., Jin, B., Brown, A. J. P., and Preece, A. (2008). Information quality in proteomics. *Brief Bioinform*, **9**(2), 174–188.

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

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., and Yaspo, M.-L. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, **321**(5891), 956–960.

Syka, J., Marto, J., Bai, D., Horning, S., Senko, M., Schwartz, J., Ueberheide, B., Garcia, B., Busby, S., Muratore, T., Shabanowitz, J., and Hunt, D. (2004a). Novel linear quadrupole ion trap/ft mass spectrometer: Performance characterization and use in the comparative analysis of histone h3 post-translational modifications. *Journal of Proteome Research*, **3**(3), 621–626.

Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004b). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences*, **101**(26), 9528–9533.

Tabb, D. L., Shah, M. B., Strader, M. B., Connell, H. M., Hettich, R. L., and Hurst, G. B. (2006). Determination of peptide and protein ion charge states by fourier transformation of isotope-resolved mass spectra. *Journal of the American Society for Mass Spectrometry*, **17**(7), 903–915.

Tan, C. S., Ploner, A., Quandt, A., Lehtiö, J., and Pawitan, Y. (2006). Finding regions of significance in seldi measurements for identifying protein biomarkers. *Bioinformatics*, **22**(12), 1515–1523.

Tang, H., Arnold, R. J., Alves, P., Xun, Z., Clemmer, D. E., Novotny, M. V., Reilly, J. P., and Radivojac, P. (2006). A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, **22**(14), e481–488.

Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005). Inspect: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal. Chem.*, **77(14)**, 4626–39.

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigo, R., Briggs, S. P., and Bafna, V. (2007). Improving gene annotation using peptide mass spectrometry. *Genome Res.*, **17**(2), 231–239.

Tanner, S., Payne, S. H., Dasari, S., Shen, Z., Wilmarth, P. A., David, L. L., Loomis, W. F., Briggs, S. P., and Bafna, V. (2008). Accurate annotation of peptide modifications through unrestrictive database search. *Journal of Proteome Research*, **7**(1), 170–181.

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q.-T. (2004). Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, **20**(17), 3034–3044.

Toll, H., Wintringer, R., Schweiger-Hufnagel, U., and Huber, C. (2005). Comparing monolithic and microparticular capillary columns for the separation and analysis of peptide mixtures by liquid chromatography-mass spectrometry. *J Sep Sci*, **28**(14), 1666–1674.

van Etten, W. C. (2006). Poisson processes and shot noise. In *Introduction to Random Signals and Noise*, pages 193–210. Wiley.

Vandenbogaert, M., Li-Thiao-Te, S., Kaltenbach, H., Zhang, R., Aittokallio, T., and Schwikowski, B. (2008). Alignment of lc-ms images, with applications to biomarker discovery and protein identification. *Proteomics*, **8**(4), 650–672.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Vardi, Y. and Zhang, C.-H. (2000). The Multivariate L1-Median and Associated Data Depth. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(4), 1423–1426.

Vincent, L. and Soille, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE PAMI, 1991*, **13**(6), 583–598.

Wagner, M., Naik, D., and Pothen, A. (2003). Protocols for disease classification from mass spectrometry data. *Proteomics*, **3**(9), 1692–1698.

Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T., Hill, L., Norton, S., Kumar, P., Anderle, M., and Becker, C. (2003). Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry*, **75**(18), 4818–4826.

Washburn, M., Wolters, D., and Yates, J. r. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**(3), 242–247.

Weiszfeld, E. (1937). Sur le point pour lequel la somme des distances de n points donnes est minimum. *Tohoku Math. Journal*, **43**, 355386.

Whistler, T., Rollin, D., and Vernon, S. (2007). A method for improving SELDI-TOF mass spectrometry data quality. *Proteome Science*, **5**(1), 14.

Whitehead, K., Kish, A., Pan, M., Kaur, A., Reiss, D. J., King, N., Hohmann, L., DiRuggiero, J., and Baliga, N. S. (2006). An integrated systems approach for understanding cellular responses to gamma radiation. *Mol Syst Biol*, **2**.

Wiley, W. C. and McLaren, I. H. (1955). Time-of-Flight Mass Spectrometer with Improved Resolution. *Review of Scientific Instruments*, **26**(12), 1150–1157.

Windig, W., Phalp, J., and Payne, A. (1996). A noise and background reduction method for component detection in liquid chromatography/mass spectrometry. *Analytical Chemistry*, **68**, 3602–3603.

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., MacInnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucl. Acids Res.*, **35**, D521–526.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **2**, 37–52.

Wolski, W. E., Farrow, M., Emde, A.-K., Lalowski, M., Lehrach, H., and Reinert, K. (2006). Analytical model of peptide mass cluster centres with applications. *Proteome Science*, **4**(18), doi:10.1186/1477–5956–4–18.

Wolters, D., Washburn, M., and Yates, J. r. (2001). An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem*, **74**, 5683–90.

Wong, J. W. H. and Downard, K. M. (2005). Performance of the computer algorithm complx for the detection of protein complexes in the mass spectra of simulated biological mixtures. *Journal of Mass Spectrometry*, **40**(9), 1187–1196.

Wu, T., Lin, C., and Weng, R. (2003). Probability estimates for multi-class classification by pairwise coupling.

Xu, M., Geer, L., Bryant, S., Roth, J., Kowalak, J., Maynard, D., and Markey, S. (2005). Assessing data quality of peptide mass spectra obtained by quadrupole ion trap mass spectrometry. *Journal of Proteome Research*, **4**(2), 300–305.

Xue, X., Wu, S., Wang, Z., Zhu, Y., and He, F. (2006). Protein probabilities in shotgun proteomics: Evaluating different estimation methods using a semi-random sampling model. *Proteomics*, **6**(23), 6134–6145.

Yates, J. r., Eng, J., McCormack, A., and D, S. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem.*, **15**, 1426–1436.

Yergey, J., Heller, D., Hansen, G., Cotter, R. J., and Fenselau, C. (1983). Isotopic distributions in mass spectra of large molecules. *Analytical Chemistry*, **55**(2), 353–356.

Yergey, J. A. (1983). A general approach to calculating isotopic distributions for mass spectrometry. *International Journal or Mass Spectrometry and Ion Physics*, **52**(2-3), 337–349.

Yu, J. S., Ongarello, S., Fiedler, R., Chen, X. W., Toffolo, G., Cobelli, C., and Trajanoski, Z. (2005). Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics*, **21**(10), 2200–2209.

Zhang, X., Asara, J., Adamec, J., Ouzzani, M., and Elmagarmid, A. K. (2005). Data pre-processing in liquid chromatography / mass spectrometry-based proteomics. *Bioinformatics*, **21**(21), 4054–4059.

Zhang, Z. and Marshall, A. G. (1998). A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Journal of the American Society for Mass Spectrometry*, **9**, 225–233.

Zhou, H., Watts, J. D., and Aebersold, R. (2001). A systematic approach to the analysis of protein phosphorylation. *Nat Biotech*, **19**(4), 375–378.

Zubarev, R. and Mann, M. (2007). On the Proper Use of Mass Accuracy in Proteomics. *Mol Cell Proteomics*, **6**(3), 377–381.

Zubarev, R., Kelleher, N., and McLafferty, F. (1998). Electron capture dissociation of multiply charged protein cations. a nonergodic process. *Journal of the American Chemical Society*, **120**(13), 3265–3266.