FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Improving the Protein Identification Performance in High-Resolution Mass Spectrometry Data

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Computational Intelligence

eingereicht von

Frederico Dusberger
Matrikelnummer 0725856

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Univ.-Prof. Dipl.-Ing. Dr.techn. Günther Raidl
Mitwirkung: Karl Mechtler
Dr. Peter Pichler

Wien, 08.10.2012

(Unterschrift Verfasser) (Unterschrift Betreuung)

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Improving the Protein Identification Performance in High-Resolution Mass Spectrometry Data

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Computational Intelligence

by

Frederico Dusberger
Registration Number 0725856

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Univ.-Prof. Dipl.-Ing. Dr.techn. Günther Raidl
Assistance: Karl Mechtler
Dr. Peter Pichler

Vienna, 08.10.2012

(Signature of Author) (Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Frederico Dusberger
Schönbrunner Straße 236/24, 1120 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasser)

Acknowledgements

I would like to thank my supervisors Günther Raidl from the Vienna University of Technology as well as Karl Mechtler and Peter Pichler from the Protein Chemistry Facility at the Vienna Research Institute of Pathology (IMP) who offered me the opportunity to work on this interesting project.

Furthermore, I am also grateful to all my colleagues at the Protein Chemistry Facility who were always helpful, encouraging and open for interesting discussions providing a productive working atmosphere. Especially, I want to thank Peter Pichler, Werner Straube and Thomas Taus for their continuous interest in my work and their critical advice and guidance.

Finally, I would like to express my gratitude to my parents Rolf and Maria Dusberger, as well as my partner Andreea Constantinescu for their constant moral support and confidence in me and my abilities during this work and my entire life.

This project would not have been possible without the help and support of any of the aforementioned persons.

Abstract

The field of proteomics is concerned with the study of structure and function of proteins. The most commonly used approach for the analysis of proteins is the bottom-up analysis where a protein is first digested into smaller peptides which are then analyzed by LC-MS/MS in order to confirm the identity of the original protein. To analyze these peptides they are first separated via liquid chromatography (LC) before their mass-over-charge ratios are recorded in the mass spectrometer as MS1-spectra. Selected peptides (*precursors*) are fragmented yielding MS2-spectra of their respective fragment ions (MS/MS). These high-throughput experiments generate vast amounts of data and are referred to as *shotgun proteomics* experiments.

For the large amount of raw data an appropriate data analysis is required in order to extract as much useful information as possible and filter out superfluous and redundant parts. However, common database search engines, which are used for identification of the peptides using their masses and the associated MS2-spectra, currently throw away most of the information contained in MS2-spectra. Moreover, the benefit of the high mass-accuracy provided by state of the art mass spectrometers is forfeited by the instruments themselves, as the MS2-spectra of the peptides are usually not recorded at the optimal time point where the intensity of the specific peptide is highest. To compensate for these drawbacks sophisticated methods are necessary that can preprocess the spectra accordingly.

In this thesis we studied the application of two ways of MS2-spectrum preprocessing to increase the number of spectra that can be identified by facilitating the identification step of the database search engine.
First, different MS2-deisotoping and -deconvolution methods were analyzed which aim for the removal of isotope peaks and peaks of multiply-charged variants of the analyte peptides. These peaks unnecessarily impair the search engine's performance by increasing the search space. We demonstrate that the algorithms raise the confidence in correct identifications by eliminating obstructing peaks, especially from the areas around correct fragment peaks. Furthermore, we show that these methods are nonetheless limited due to the design of the scoring algorithms of common search engines.
Secondly, to fully exploit the information that is made available through high mass-accuracy, we developed a 3d-peak picking algorithm that does not rely on the peptide mass information of the single MS1-spectrum it was selected from for fragmentation but additionally reconstructs the peptide's elution profile gathering many data points to obtain a statistically confident value for the mass. Experiments demonstrated that peptide masses calculated from reconstructed 3d-peaks have a significantly higher precision than using the conventional precursor mass values

provided by the instrument. We show that the high precision also increases the identification performance, especially for strict search tolerances.

The designed algorithms were implemented in a plugin for a commercially available software package (Proteome Discoverer by Thermo Fisher Scientific) which is now used in the proteomics group of Karl Mechtler. Moreover, the plugin is available for download, free of charge.

Kurzfassung

Die Proteomik befasst sich mit der Struktur und Funktion von Proteinen. Der am weitesten verbreitete Ansatz zur Analyse von Proteinen ist die “bottom-up”-Analyse, bei der ein Protein zuerst in kleinere Peptide verdaut wird, welche dann mittels LC-MS/MS analysiert werden, um die Identität des ursprünglichen Proteins zu bestätigen. Um diese Peptide analysieren zu können, werden sie zunächst mittels Flüssigchromatographie (LC) aufgetrennt. Anschließend wird deren Masse-zu-Ladung-Verhältnis im Massenspektrometer gemessen und in Form von MS1-Spektren aufgezeichnet. Ausgewählte Peptide (*Precursor*) werden fragmentiert, was zu MS2-Spektren der jeweiligen Fragmentionen führt (MS/MS). Diese Hochdurchsatzexperimente erzeugen immense Datenmengen und werden als *Shotgun Protemoics*-Experimente bezeichnet.

Diese große Menge an Rohdaten muss durch adäquate Methoden analysiert werden, um so viele nützliche Informationen, wie möglich zu extrahieren und überflüssige, redundante Teile herauszufiltern. Die verbreiteten Datenbank-Suchmaschinen, die zur Identifikation der Peptide mittels ihrer Masse und der zugehörigen MS2-Spektren herangezogen werden, verwerfen zur Zeit einen Großteil der Informationen im MS2-Spektrum. Zudem wird der Vorteil der hohen Massengenauigkeit, welche mit den modernsten Massenspektrometern erreichbar ist, durch die Geräte selbst wieder eingebüßt. Dies hat den Grund, dass die MS2-Spektren der Peptide in der Regel nicht zum optimalen Zeitpunkt, zu dem die Intensität des Peptids am größten ist, aufgenommen werden. Um diesen Nachteilen entgegenzuwirken, sind ausgefeilte Methoden für entsprechendes Preprocessing der Spektren nötig.

In dieser Diplomarbeit untersuchen wird zwei Arten von Preprocessing-Methoden für MS2-Spektren, mit dem Ziel die Anzahl der Spektren, die identifiziert werden können zu erhöhen, indem der Identifizierungsprozess, der von der Datenbanksuchmaschine durchgeführt wird, vereinfacht wird.
Erstens werden verschiedene MS2-Deisotoping und -Deconvolution Methoden untersucht, welche das Ziel haben, Isotopen-Peaks und Peaks mehrfach geladener Varianten der Analytpeptide zu entfernen. Durch die Vergrößerung des Suchraums beeinträchtigen diese Peaks unnötigerweise die Leistung der Suchmaschine. Wir führen aus, dass die Algorithmen das Vertrauen in die korrekte Identifikation von Peptiden durch das Entfernen von Peaks, vor allem aus den Bereichen um die korrekten Fragment-Peaks, welche andernfalls das Finden dieser korrekten Peaks erschweren würden, erhöht. Außerdem zeigen wir, dass diese Methoden nichtsdestotrotz durch das Design der Scoring-Algorithmen verbreiteter Suchmaschinen eingeschränkt sind.
Zweitens entwickeln wir einen 3d-Peak-Picking Algorithmus, der sich im Bezug auf die Masse der Peptide nicht allein auf die Infomation des einzelnen MS1-Spektrums verlässt, aus welchem

das Peptid zur Fragmentierung ausgewählt wurde. Es wird statt dessen zusätzlich das Elutionsprofil des Peptids rekonstruiert, wobei viele Datenpaunkte erfasst werden, um einen statistisch zuverlässigen Wert für die Masse zu erhalten. Somit ist es möglich die Information, die durch die hohe Massengenauigkeit erreichbar ist voll und ganz zu nutzen. Unsere Experimente zeigen, dass die Peptidmassen, welche aus den rekonstruierten 3d-Peaks berechnet wurden, im Vergleich zu den vom Gerät zur Verfügung gestellten Massen, eine wesentlich höhere Präzision besitzen. Darauf aufbauend zeigen wir zudem, dass diese hohe Präzision die Anzahl der identifizierten Peptide, vor allem für strenge Suchtoleranzen, steigert.

Aus den entwickelten Algorithmen ist ein Plugin für ein kommerziell verfügbares Softwarepaket (Proteome Discoverer von Thermo Fisher Scientific) entstanden, welches nun in der Proteomikgruppe von Karl Mechtler eingesetzt wird. Zudem ist dieses Plugin kostenlos zum Download verfügbar.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Problem definition and aim of the work . . . 2
- 1.3 Related Work . . . 3
- 1.4 Outline of the Thesis . . . 4

2 Proteomics and Mass Spectrometry **5**
- 2.1 Proteomics . . . 5
  - 2.1.1 Amino Acids, Peptides and Proteins . . . 5
- 2.2 Mass Spectrometry . . . 6
  - 2.2.1 Composition of a mass spectrometer . . . 8
  - 2.2.2 The Shotgun-Proteomics approach . . . 12
    - 2.2.2.1 Proteolytic digestion of large proteins into peptides . . . 13
    - 2.2.2.2 Peptide separation by High-Performance Liquid Chromatography . . . 14
  - 2.2.3 Peptide ionization in the ion source . . . 15
  - 2.2.4 Tandem Mass Spectrometry . . . 16
    - 2.2.4.1 Fragmentation of peptide ions by activation leading to MS/MS-spectra . . . 17
  - 2.2.5 Analysis of generated mass spectrometry data . . . 19
    - 2.2.5.1 Spectrum Preprocessing . . . 19
    - 2.2.5.2 Database Search . . . 19
    - 2.2.5.3 Evaluation of search results . . . 21
- 2.3 Definition of relevant data structures . . . 25
  - 2.3.1 The .RAW file format . . . 25
  - 2.3.2 Formal definition of mass spectra . . . 26

3 MS2 - Spectrum Manipulation **29**
- 3.1 Motivation . . . 29
- 3.2 Deisotoping . . . 29
  - 3.2.1 A simple Deisotoping algorithm . . . 31
  - 3.2.2 Improving the original algorithm . . . 34
    - 3.2.2.1 Expected isotope-ratio determination by averagine modeling 35

3.2.2.2 Reasons for the superiority of the simple approach 39
3.3 Charge-Deconvolution of spectra 41
3.3.1 A simple Charge-Deconvolution algorithm 42
3.4 Results and Future Work 43

4 Precursor Precision Improvement **47**
4.1 Motivation 47
4.2 Method 48
4.2.1 Basic Principle: 3d-Peaks 48
4.2.1.1 2d-Peak Detection 48
4.2.1.2 3d-Peak Centroid 52
4.3 3d-Peak Reconstruction 52
4.3.1 Minimum Profile Points in 2d-Peak 55
4.3.2 Peak to Peak m/z-Deviation 56
4.3.3 Continuous m/z-Deviation 56
4.3.4 Elution Profile Tracking 57
4.3.5 Evaluation of the extracted peak 59
4.4 Further ways of improving the mass precision 60
4.4.1 Usage of Isotope Patterns 60
4.4.2 Linear Mass Recalibration 61
4.5 Resulting Software 62

5 Experimental Results **67**
5.1 Experimental Setup 67
5.2 MS2 - Spectrum Manipulation Results 68
5.3 Precursor Precision Improvement Results 68

6 Conclusion and Future Work **85**

List of Figures **87**

Bibliography **89**

CHAPTER 1

Introduction

1.1 Motivation

Mass spectrometry, in particular tandem mass spectrometry, has become one of the standard approaches for analyzing proteins and peptides. The common approach is the so-called bottom-up proteomics which is executed by employing a high-throughput liquid chromatography coupled to tandem mass spectrometry setup (LC-MS/MS). Because of the immense number of proteins, both in the biologic samples, as well as in the databases that are used for identification, efficient algorithms are necessary to cope with the amount of data.
Another aspect making the process more difficult is the fact that only a subset of the peaks in the spectra produced by the mass spectrometer contains relevant information which can be used for searching in a protein database. This is the case because a part of each spectrum is simply noise caused by the inaccuracy of the instruments (*electronic noise*). On the other hand, there is always chemical contamination (*chemical noise*) caused, for example, by polymers or charged ions occurring in the air. From the viewpoint of a protein search engine, the resulting peaks cannot be used for identifying a peptide ion and are therefore also to be seen as a kind of noise. Furthermore, considering the MS2-spectra, ions with a similar m/z-value in the isolation window, which consequently will be fragmented at the same time, can lead to so called chimera spectra. Because of this a spectrum can thus also contain even more peaks which are, regarding the identification of the actual precursor, to be considered as noise.
Also the part of a spectrum, in which distinct peaks stemming from the correct fragments indicate the precursor ion which is to be identified, is not completely evaluated by the algorithms that have been developed, so far: Peaks of ^{13}C-isotopes, as well as multiply-charged occurrences of one and the same fragment ions leading to predictable additional peaks in the spectrum are not sufficiently recognized or filtered by many programs. The additional confirmation that recognizing these situations would contribute to the identification of a peptide is therefore forfeited, as well.

Furthermore, most of the algorithms developed for noise reduction that are currently being used have been developed especially for older instruments which are, regarding the resolution

and the accuracy of the measurements, by far inferior to the instruments that exist nowadays. Therefore, these algorithms are, of course, not optimized with respect to the current state of the art mass spectrometers.
One crucial advantage of most recent instrument generation is thus that the MS2-spectra output by them suffer far less from electronic noise compared to older instruments. Noise reduction in the classical sense, i.e. the very general removal of noise via filtering algorithms, is therefore less useful. Moreover, as most of the peaks in an MS2-spectrum are correct (non-noise) peaks such a procedure would rather be counter-productive as it would remove mostly these correct peaks.

Therefore, the opportunity for developing better algorithms for high-resolution spectra should rather be sought in the well-directed removal of the above-mentioned isotope peaks and the recognition of ions that occur in multiple charge states, as well as in the proper usage of this additional information.
In addition to that, high-resolution instruments offer many more possibilities for improvement due to the accurate determination of ion masses. One of these is, for example, the recomputation of the precursor mass from one or more adjacent MS1-spectra. By taking not only the monoisotopic precursor peak from the respective MS1-spectrum but also possible isotope peaks of the precursor ion into account, one could obtain a more precise value for its mass. This idea is applicable even better, if peaks indicating the same precursor are gathered from different MS1-spectra, as well, increasing the number of data points that can be used. Moreover, further processing methods can be applied that aim for the detection of certain instrument-specific shifts within the measured data. Such an existing instrument bias can be removed by performing a recalibration of the data using further processing algorithms.

1.2 Problem definition and aim of the work

Protein database search engines employ scoring schemes in order to express the confidence in the explanation of a certain spectrum by the existence of a certain ion peptide in the sample. For example, the search engine Mascot [35] bases its scoring upon the probability that the found explanation for the fragment ions in the MS2-spectrum is just mere coincidence. The smaller this probability, the higher the score Mascot assigns to the matched peptide.

The problem that we want to address in this thesis can now be stated quite succinctly. Due to the several above-mentioned factors, such as electronic and chemical noise, the measurement error of the instrument, miscalibration, etc., protein identification is far from being perfect, i.e. there is still a considerable amount of spectra which remains unexplained. The reason is that the search engine cannot assign a matching peptide to the spectrum with a probability that is high enough to accept it as a confident identification.

The aim of this thesis is therefore to develop and test new algorithms that are designed for the application on high-accuracy, high-resolution spectra, as are generated by modern mass spectrometers. By processing the MS1- and MS2-spectra in an intelligent way, we want to achieve an improvement of the identification rate by eliminating or at least reducing the influence of some of these perturbing factors. More precisely, this means an increase of scores for correct identifications and a decrease of the scores assigned to wrong explanations. As pointed out

above, different methods than the currently established ones are necessary to achieve this goal, but on the other hand the higher accuracy also opens up new possibilities of gathering more useful information from these spectra than from the ones generated by older instruments having lower resolutions.
In general, the algorithms we want to develop are based on the following three concepts:

- Gain information from the spectra which can be used for further pre- (or post-) processing steps.
  This aspect focuses on finding certain properties of the fragment ions and the precursor which can then be applied in an additional step to modify the spectra accordingly.

- Gain information from the spectra that helps setting specific search parameters for the search engine.
  Here the relevant data extracted from the spectra directly influences the search, e.g setting the mass tolerance or the search interval, etc.

- Alter MS2-spectra and the associated metadata directly in order to influence the identification process. This includes method that change the m/z-values or the intensities of the peaks stored in the spectra. Furthermore, data attached to the spectrum, such as parts of the precursor information can be modified.

In the end, the new algorithms should then be integrated into the Proteome Discoverer Software (Thermo Fisher Scientific, version 1.3.0.339) where users can choose to integrate them into their workflows in order to apply them as additional preprocessing steps.

1.3 Related Work

After the development of modern mass spectrometry which allowed for large-scale analysis of the proteome [1] tandem mass spectrometry coupled to liquid chromatography (*LC-MS/MS*) is nowadays the method of choice for analyzing the proteins contained in complex samples [33,47]. For this reason, of course, various spectrum preprocessing approaches (for MS1-, as well as MS2-spectra) have been developed with the aim of improving the results achieved with the LC-MS/MS method. However, many of these are based upon older instruments and employ, for instance, Fourier transform which is suited to clean spectra containing less sharply-defined peaks [26,27]. Other authors follow statistical approaches trying to compare the spectrum to the theoretical model spectra of known peptides in order to determine the one that provides the best match [11,24,40].
There are also approaches which could make use of the higher mass accuracy of modern instruments. There, the *centroid peaks*, i.e. peaks being determined by the centroid of the ion's measured distribution, are directly used, since they are distinct enough for allowing comparisons between measured distances between these peaks and the theoretically expected distances [38].
Finally, there are also methods where the information gained from recognizing correlations in the spectra is indeed used. This is done by applying an own scoring scheme [38] or even combining several different scoring schemes [28] for the evaluation of the possible candidates for

identification. The problem, however, is that the scoring can only be applied as a postprocessing filter as the protein database search engines use their own scoring system, which can hardly be influenced by external settings.

Of course, there is also a number of approaches that do not modify the spectra per se, but try to improve the measuring method following the goal of increasing the mass precision of the measured precursor ions. Recently, Zhang et al. described a way of dynamically choosing one or more lock masses for recalibration of MS1 spectra instead of using a fixed lock mass [51].

Regarding the acquisition of additional information about the precursor ion, especially the determination of the elution profile of the respective ion is an established method. The search engine MaxQuant [7] relies heavily on this principle, which allows for highly accurate precursor masses leading to high peptide identification rates. This software suite has also been equipped with an algorithm to perform a recalibration of the precursor masses according to the results obtained from a first search. The new masses are than used for the second, actual search [8].

1.4 Outline of the Thesis

The first chapter of this thesis has provided an introduction and a short overview of the topic. Moreover the problem the thesis is concerned with has been defined and the aims have been lined out.
In the following chapter a more thorough introduction to the field of proteomics and to mass spectrometry in particular will be given, providing the relevant background information. Furthermore, technical terms, which are used in the following chapters of the thesis are explained and put into context.
The third chapter deals with MS2-spectrum manipulation and explains the need for a deisotoping and a charge-deconvolution procedure for MS2-spectra. In both cases the context is explained and the necessary theoretical background knowledge is provided before the existing algorithm which was used as a basis is explained. Afterwards the algorithm is assessed and remedies for possible weaknesses are developed. Subsequently, additional ideas on MS2-spectrum manipulation are sketched. They were, however, not implemented as this would have gone beyond the scope of the thesis. The chapter concludes with a short description of the resulting final software product.
Chapter 4 provides a detailed step-by-step explanation of the work on the precursor mass-precision improvement. For each part of the algorithm, the underlying idea is described and accounted for with an example. It also introduces a concept for recalibration of precursor masses according to high-confident search results of a preliminary search. Finally, as in the previous chapter, a summary of the resulting software products is given.
In chapter 5 a description of the instances with which the algorithms were tested, i.e. the relevant sample properties and instrument settings. The experimental results of the algorithms are then listed along with the corresponding figures. Finally, chapter 6 summarizes the results and discusses possible future work.

CHAPTER 2

Proteomics and Mass Spectrometry

2.1 Proteomics

The field of proteomics deals with the study of structure and function of proteins, i.e the study of the *proteome*, which can be seen as the functional representation of the genome [5]. The term proteome was coined by Mark Wilkins [48] and is derived from proteins. It defines the entirety of proteins expressed by the genome. While the genes in a genome remain static the gene products, i.e. the *proteins* that are actually present in a cell are, in general, different given different biological contexts. An example showing the difference between genome and proteome and illustrating how strong the effect of changes in the proteome can be is the metamorphosis of a caterpillar into a butterfly. The caterpillar and the butterfly have the very same genome. However, through the expression of different genes and the resulting difference in the proteome, the organism changes completely, which we can see best in its appearance.
Another contribution to the high dynamics of the proteome is that most of the proteins are subject to chemical modification, also called *post-translational modifications* or PTMs.

2.1.1 Amino Acids, Peptides and Proteins

A protein is a linear polymer ("a sequence") of monomer building blocks called *amino acids*. These are molecules consisting of a central carbon atom, linked to an amino group (NH_2), an acidic carboxylic group (COOH), a hydrogen atom (H) and a side chain (or simply *R* group). This side-chain is what defines an amino acid, as it is composed of a different group of atoms for each amino acid being the only difference among them. It is also called the functional group of the amino acid. Figure 2.1 depicts this generic structure of an amino acid. An amino acid usually occur ionized depending on their environment, i.e. the pH and other molecules it might be linked to. This means that the amino group can be protonated, which leads to a one-fold positively charged NH_3^+ and the carboxyl group can be deprotonated leading to a one-fold negatively charged COO^-. There are 20 amino acids, the so-called proteinogenic amino acids, that can be naturally incorporated into proteins.

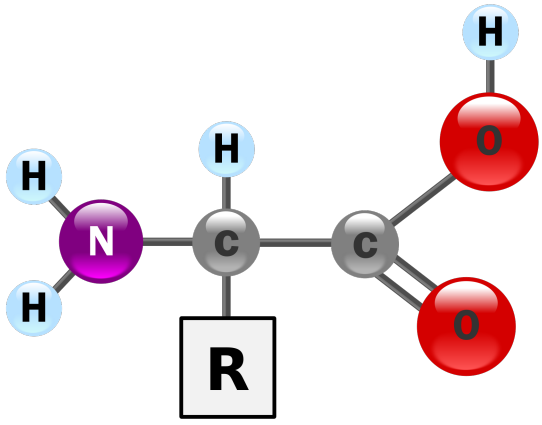


Figure 2.1: The generic structure of an amino acid in its unionized form. It consists of the central C-atom which is linked to the amine group (NH_2), the carboxyl group (COOH) a Hydrogen and its functional group (R) via covalent chemical bonds [45].

One amino acid having a protonated amino group and another one with a deprotonated carboxyl group can now be linked together to a dipeptide[1] via a so-called *peptide bonds*. These are covalent chemical bonds whose formation is accompanied by the loss of a water molecule, as can be seen in the illustration of a peptide bond link in Figure 2.2. This reaction is also called condensation. A series of amino acids linked by a peptide bond is called a *polypeptide*. Proteins have unique amino acid sequences, which are referred to as the *primary structure* of a protein. The two ends of such a sequence are terminated by a free carboxyl group (COOH) and a free amine group (NH_2) and are referred to as *C-terminus* and *N-terminus*, respectively. Figure 2.3 shows a peptide consisting of four amino acids with highlighted C- and N-terminal ends.
Based upon this primary structure which is illustrated in Figure 2.4, one can study the secondary, tertiary and quaternary structure of proteins, which emerge as the molecules strive to achieve the energetically most convenient arrangement in space and therefore fold itself in determined ways. An overview on these different structure levels can be seen in Figure 2.5.

2.2 Mass Spectrometry

As proteins play a crucial role in nearly all biological processes, the understanding of their functions, structure and dynamics is of major interest in biochemistry. However, when analyzing biological samples one has to cope with the vast number of different proteins expressed in a cell. In this field *mass spectrometry* has become a widely used method for the large-scale analysis of complex protein samples. It replaced the *Edman degradation* [12], which has previously been considered the method of choice in protein identification. Nowadays, however, the Edman degradation is considered too slow and inaccurate for large-scale protein identification. [34]. Another

[1] Chemically speaking there is no difference between a peptide and a protein. It is a common convention to refer to proteins having less than 50 amino acids as peptides.

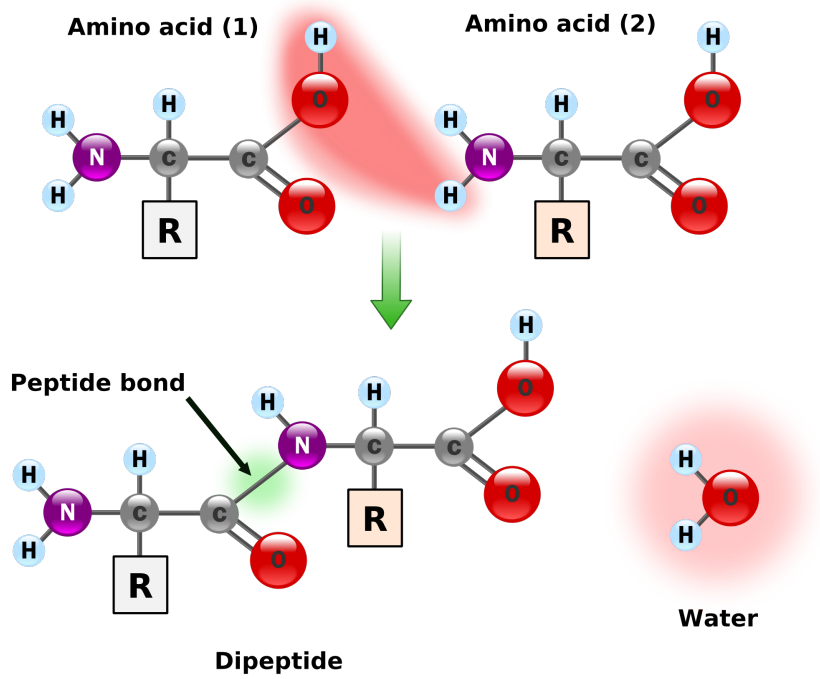


Figure 2.2: Two amino acids forming a dipeptide via a peptide bond. A condensation reaction takes place during which a water molecule is split off and the remaining CO^{+}- and NH^{-}-groups are linked together [45].

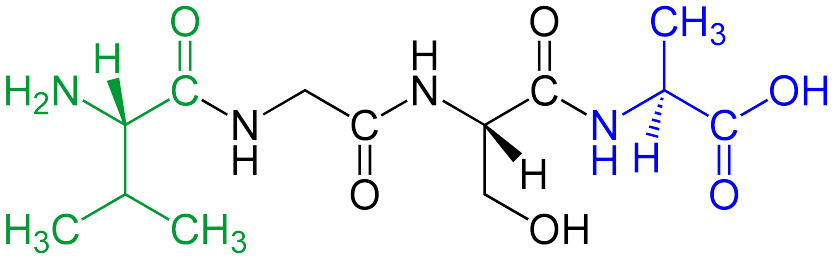


Figure 2.3: A peptide (example: Val-Gly-Ser-Ala) with green highlighted N-terminal amino acid and blue highlighted C-terminal amino acid [45].

technique commonly used in protein identification is the so-called *Western Blot* [46]. However, this approach only detects the presence of proteins that are targeted by specific antibodies. Mass spectrometry on the other hand, provides the opportunity to obtain a more general picture as it operates unspecifically and can in theory detect every protein that is abundant enough in the cell. It therefore fills the gap left by the Edman degradation overcoming the challenges of complex samples.

The basic idea behind mass spectrometry is the determination of the mass-to-charge ratio (m/z) of ionized peptides and proteins in the gas phase, which is measured in *Thompson*. With *Dalton*

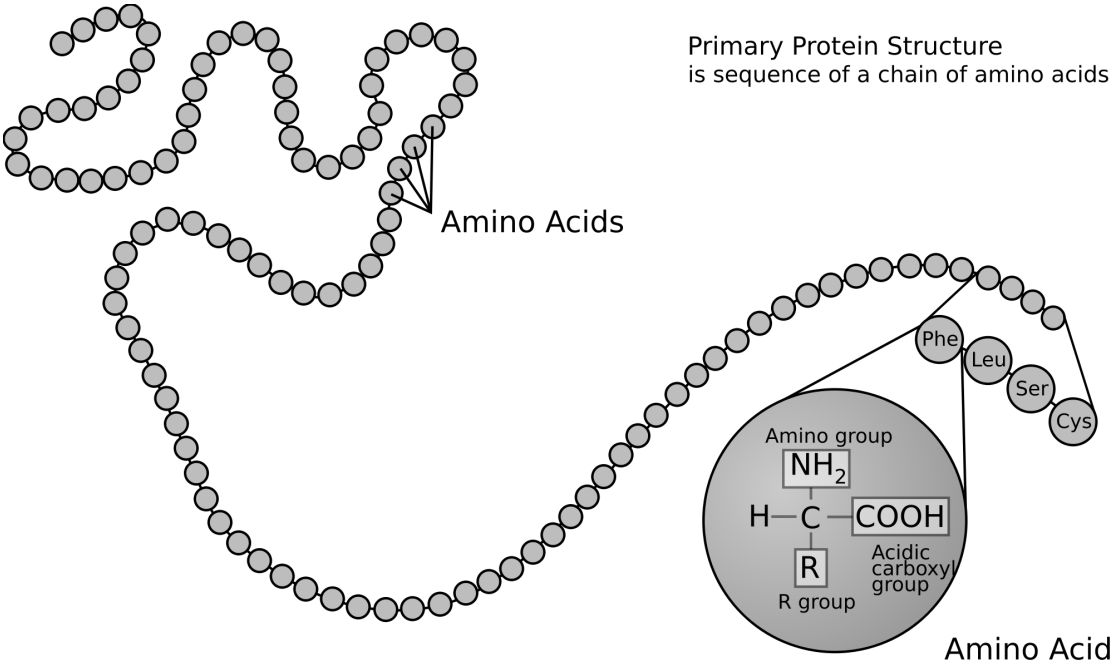


Figure 2.4: The primary structure of a protein, a sequence of amino acids linked by peptide bonds [45].

(abbreviated by *Da*) being the unit in which the atomic masses are measured and the number of charges z, which is dimensionless, it is a generally accepted convention that the m/z-value is, simply, also given in Dalton. For convenience one also talks about *mass-units* (*mu*) and *milli-mass-units* (*mmu*), where 1 mu is simply the same as 1 Da and consequently 1 mmu is equal to 0.001 Da.
Another common unit, especially used when talking about the mass deviation (or the deviation in m/z) between two ions, is *parts-per-million* (abbreviated with *ppm*). Moreover, it is a common means of defining the accuracy of measurements. When an ion is measured, the error in ppm between the observed m/z-value m_O and the theoretical value m_T is given by

$$\frac{m_O - m_T}{m_T} \cdot 10^6 \tag{2.1}$$

2.2.1 Composition of a mass spectrometer

Mass spectrometers generally consist of three major components, an ion source, a mass analyzer and a detector. Figures 2.6 and 2.7 show a simplified scheme of a mass spectrometer and an overview of an actual instrument, the Thermo Fisher Scientific Q-Exactive mass spectrometer, respectively.
Ion source:

The analysis of a sample starts with the introduction of the sample into the ion source, where the

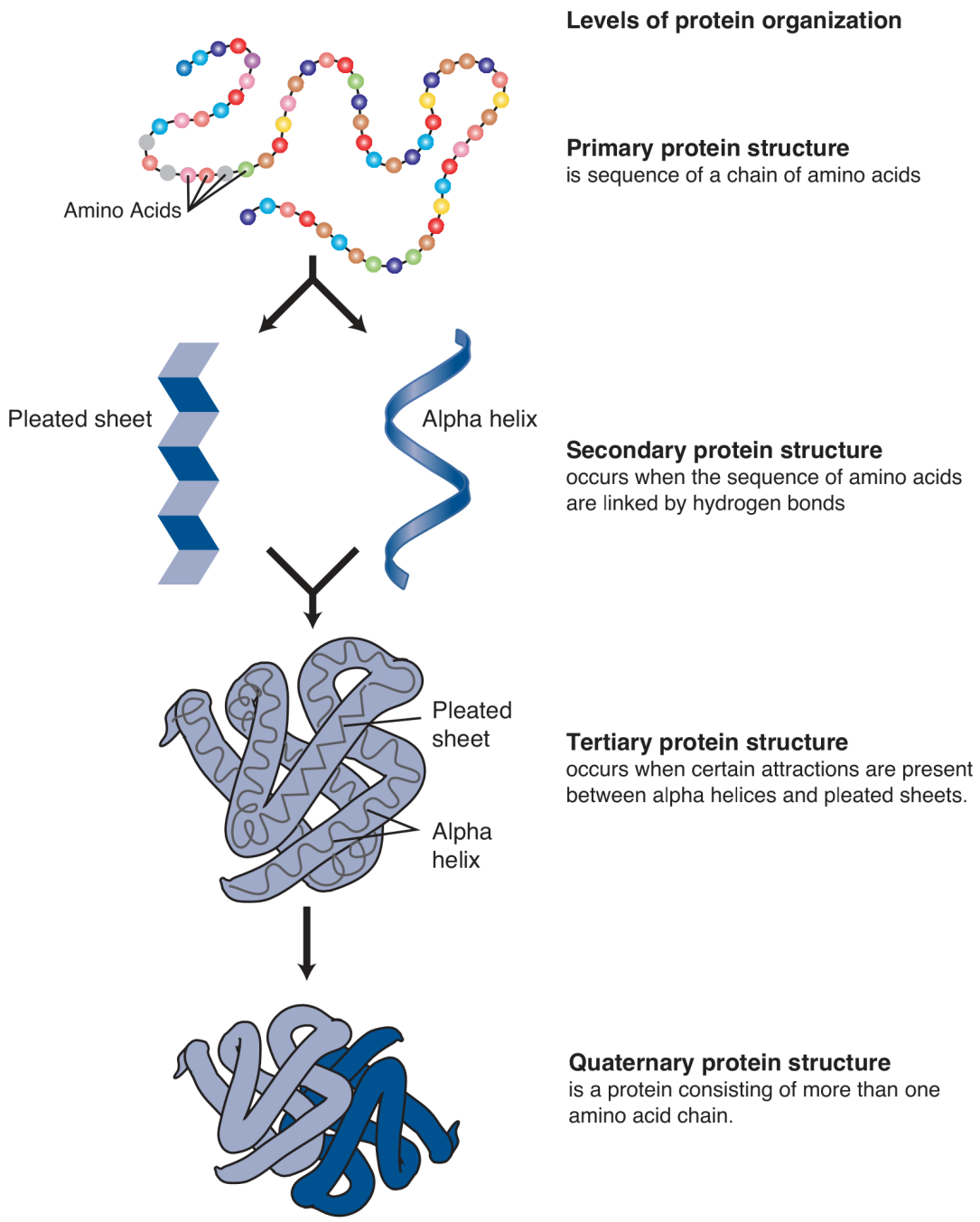


Figure 2.5: The relationship between the different structure levels of a protein [30].

analyte molecules are ionized. Most commonly, this is done by *Electrospray ionization (ESI)*

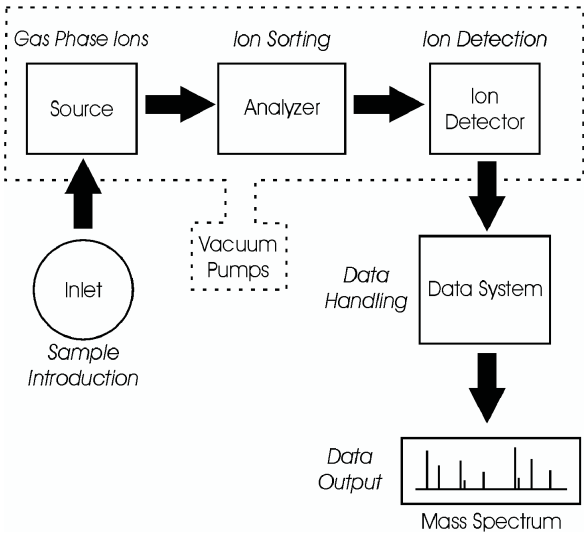


Figure 2.6: A mass spectrometer is generally composed of three parts: First the analyte molecules are ionized in the *ion source*. The *mass analyzer* then determines the m/z-ratios of their relative abundance is measured in the *ion detector*. Subsequently the generated data is processed and analyzed [26].

or *matrix-assisted laser desorption/ionization (MALDI)*, which are the two major techniques for the ionization of proteins and peptides.

Mass analyzer:

Then, the mass analyzer separates the incoming ions according to their m/z-ratio. Four different kinds of mass analyzers currently used in proteomics can be distinguished [1]

1. *Ion trap*:
   Ions of a specific mass range are collected (*trapped*) for a certain time by the use of a quadrupolar radio-frequency electric field. Ions of a wide m/z-range are confined within the trap and according to the applied field only ions with a certain m/z-ratio are expelled [10]. When the trap is full or a certain target threshold is reached the ions can be released by gradually changing the frequency in order to undergo the further MS1 or MS2 analysis. The ion trap's strengths are robustness and sensitivity, however it only provides a relatively low mass accuracy.

2. *Time-of-flight* (TOF):
   This instrument determines the m/z-ratio of ions by measuring their time of flight. The essential idea is that the ions, which are - driven by an electric field - moving in the same direction through a tube, have a (more or less) constant kinetic energy. As the velocity of an ion is inversely proportional to the square root of its m/z-ratio, ions having different m/z-ratios will arrive at the target plane at different times [17].

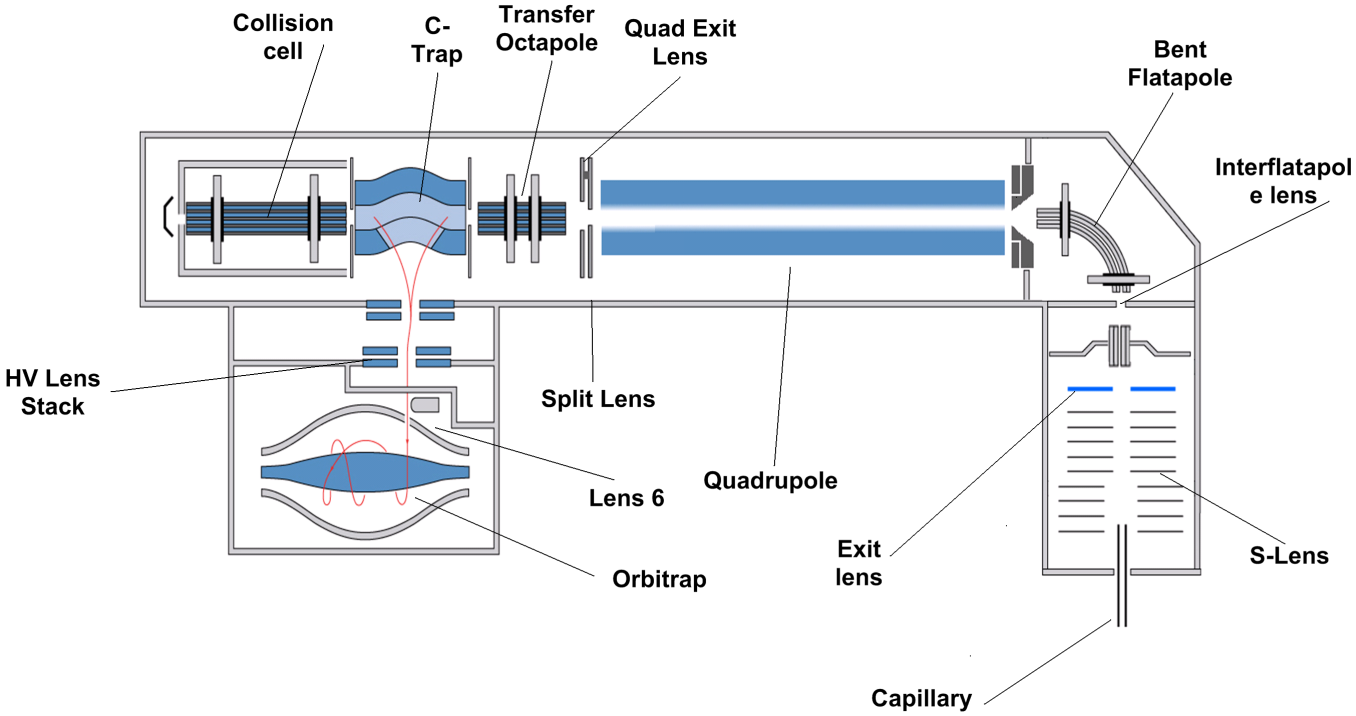


Figure 2.7: Scheme of a QExactive Orbitrap instrument [Thermo Fisher Scientific]

3. *Quadrupole*:
   A quadrupole consists of 4 circular - ideally, hyperbolic - metal rods which are aligned in parallel to each other. Between these rods a quadrupolar alternative electric field is superposed on a constant field. The key idea is that according to the applied voltage only ions with a certain m/z-ratio will pass through the quadrupole and reach the detector. Ions having a different ratio have unstable trajectories and will eventually hit one of the rods [10].

4. *Fourier transform ion cyclotron* (FT-MS):
   This type of mass analyzer basically also traps desired ions. However, it does so by applying strong magnetic fields. It provides a high sensitivity, mass accuracy, resolution and dynamic range. Despite these many advantages it is not widely used in proteomics research because of its high expense, operational complexity and especially its low peptide-fragmentation efficiency.

The mass analyzer can be seen as the key element of the mass spectrometer, since its task is the actual selection and filtering of ions, which is necessary to dsicrimante between them.

Ion detector:

Finally, the separated ions arrive in the ion detector where the relative abundance of ions for each m/z-value (within the bounds of the resolution) is registered.
As for the mass analyzer, there are several different ion detectors that are commonly used:

1. *Faraday Cup*:
   The Faraday Cup employs a very simple principle. The incoming ions hit its metal surface which leads to the emission of electrons (so-called *secondary emission*) that induce a current. This current is then amplified and recorded.

2. *Electron Multiplier*:
   The electron multiplier extends the principle of a Faraday Cup. It consists of a vacuum tube containing a series of metal plates which are maintained at increasing electrical potentials. Incoming ions hit the first plate leading to the emission of electrons. These are attracted by the second plate which in turn emits electrons upon being hit. This way, a cascade of secondary emissions is triggered at whose end typical amplification rates in the order of $1{:}1^6$ are achieved.

3. *Orbitrap*:
   The orbitrap consists of two electrodes, an inner spindle-like one and an outer electrode which envelops the inner one. Ions are injected tangentially into the orbitrap and trapped inside by the balance between the electrostatic attraction to the inner electrode and the centrifugal force. The ions therefore cycle around the inner electrode forming rings which oscilate along the electrode. This oscilation process is inversely proportional to the square root of the mass-to-charge ratio and can therefore be used to detect abundant ions.

2.2.2 The Shotgun-Proteomics approach

The term *Shotgun Proteomics* refers to a common form of non-targeted bottom-up proteomics to analyze complex protein mixtures [42]. It is non-targeted because it does not aim to detect a specific protein but rather tries to identify as many proteins as possible by comparison against possible theoretical matches in a database. The term bottom-up accounts for the prior digestion of the proteins into smaller peptides before running the MS/MS analysis.
A shotgun proteomics workflow (see Figure 2.8) consists of the following steps:

- Digestion (*Cleaving*) of the proteins into smaller peptides by an enzyme (e.g trypsin).
- Separation of the peptides by liquid chromatography or *high-performance liquid chromatography (HPLC)*.
- Ionization of the peptides by *electrospray ionization (ESI)* or *matrix-assisted laser desorption/ionization (MALDI)*.
- Analysis of the peptide ions by tandem mass-spectrometry (MS/MS).
- Comparison of the generated data against databases in order to identify the proteins present in the sample.

In the following these steps are described in detail.

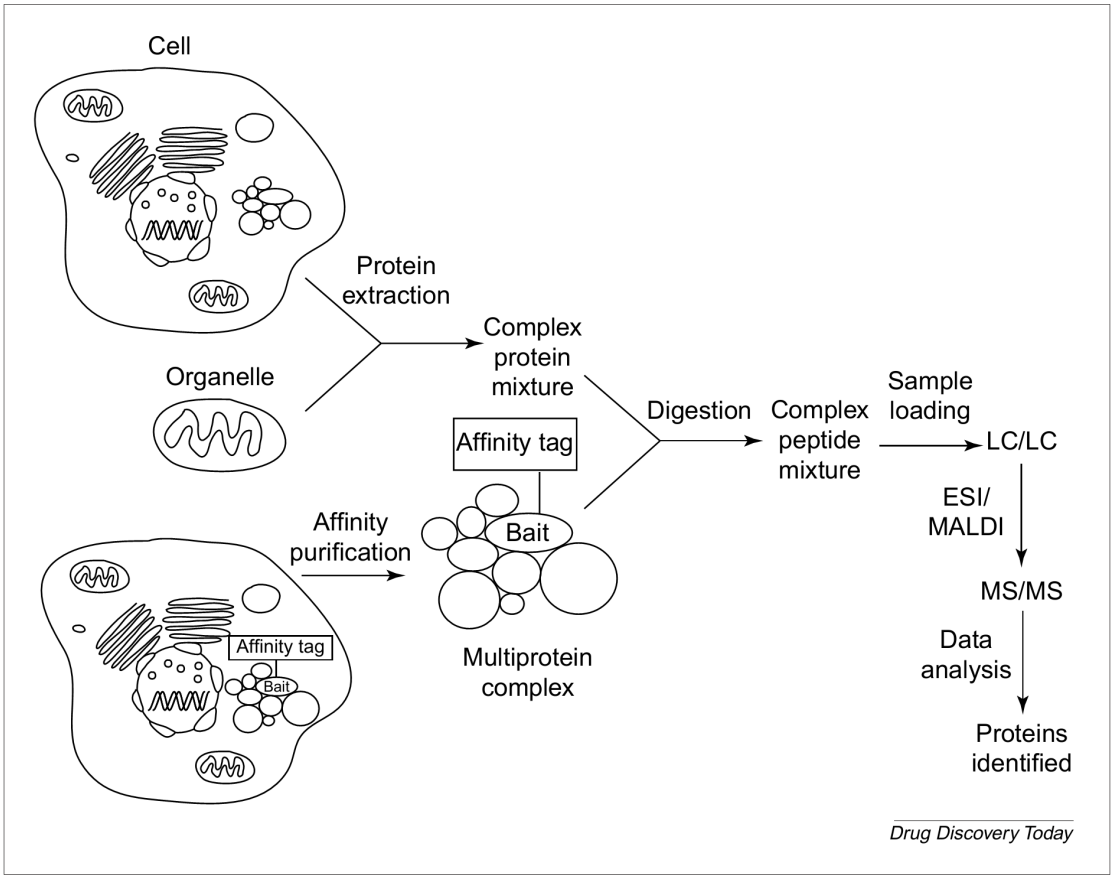


Figure 2.8: Typical workflow of a shotgun proteomics experiment: First, the proteins are extracted from the biological sample. Depending on the sample the resulting protein mixture can become very compley in terms of different proteins which are abundant in the mixture. The proteins are then cleaved into smaller peptides by proteolytic digestion leading to a more complex peptide mixture. To reduce the complexity the peptides are seaparated by (multi-dimensional) LC and ionized using ionization techniques such as ESI or MALDI. These ions are then subjected to an MS/MS analysis. Finally, the analysis results are processed in order to determine the abundant proteins [42].

2.2.2.1 Proteolytic digestion of large proteins into peptides

Proteolytic digestion is an important step to facilitate the analysis of the proteins cleaving them into smaller peptides. This approach has several advantages: On the one hand, the ionization of whole proteins leads to a great variety of charge states, whereas peptides are mostly two-fold or three-fold charged. Additionally, whole protein ions lead to huge isotope clusters in the resulting spectra, which further complicates their analysis, whereas peptides are usually small enough, such that they usually only yield clusters containing around one to four isotopes. The reason is simply that the smaller a molecule is, the less atoms exist in it that could possibly

be occurring not as the usual, most likely isotope variant but as a heavier one. Finally, the identification of several peptides that can originate from the cleaving of a specific protein is a much stronger proof for the correctness of the protein-identification than simply finding a match for the whole protein.

Thus, the initial step of the workflow is the proteolytic digestion, i.e. the cleaving of a protein into smaller peptides by using a suitable endoproteolytic enzyme. These enzymes break peptide bonds at specific positions in the peptide, i.e. before or after certain nonterminal amino acids. Unfortunately, this proteolytic digest is not always complete and it is possible that some of the specific cleavage sites are missed (so-called *missed cleavages*). Trypsin has proved to be a suitable choice for this purpose, as it has been showed that Trypsin cleaves exclusively C-terminal to Arginine and Lysine residues. [32]. Having the highly basic amino acids Lysin and Arginine at the C-termini of the peptides also facilitates the ionization of these peptides, since they have a high affinity to protons. Another advantage is that a tryptic digest leads to peptides whose masses are in the preferred range for effective fragmentation by a subsequent MS/MS analysis, which is due to the fact that Arginine and Lysine each account for around 5% of the amino acids in a protein. Since there are 20 proteinogenic amino acids the protein is in theory cleaved after every 10-th amino acid, i.e. the average length of the resulting peptide is ca. 10 amino acids. These factors have made Trypsin the most common choice when conducting a shotgun proteomics experiment.

2.2.2.2 Peptide separation by High-Performance Liquid Chromatography

The number of distinct proteins in a complex sample is too high to allow for identification using only a handful of scans. Moreover, the complexity of the sample is further increased by the digestion of the protein mixture. Therefore, the peptides obtained after the digestion step must be separated, such that ideally only one peptide at a time is analyzed by the mass spectrometer. In general, there are several key characteristics according to which the peptides in the mixture can be separated: solubility, size, charge and hydrophobicity. The technique of choice generally used in shotgun proteomics is the *reversed-phase high-performance liquid chromatography* [1], which makes use of the hydrophobicity. In general, liquid chromatography employs two phases, a solid stationary one (also called the *column*) and a liquid one. The column contains certain chemical groups to which the peptides have a greater or lesser affinity, which causes them to pass the column more or less quickly. In reverse-phase liquid-chromatography the solid phase contains hydrophobic molecules whereas the liquid phase is highly hydrophilic. Consequently, the more hydrophilic a peptide is, the quicker it will pass through the column. In contrast, hydrophobic peptides will get caught more often on their way by intermolecular interactions with the stationary phase, slowing down their progress. In order to accelerate the elution of highly hydrophobic peptides, which would bind to the column for far too long, concentration gradients of organic compounds are used, i.e. the hydrophobicity of the liquid phase is gradually increased. The use of a gradient is opposed to the so-called isocratic elution methods, where the concentration of organic compounds remains constant through the whole elution process. Typically, the liquid phase contains 0.1 % formic acid and a gradient of 5 to 75 % acetonitrile is used. The time a peptide takes through the column from the moment it enters to the moment it elutes is referred to as its *retention time*.

The *high-performance* component of the name originates from the much higher pressures that needs to be applied to the column in order to maintain a constant flow-rate. This is necessary as the column is made of much finer material than *normal* chromatography columns. This results in higher resolutions through better separation.

To further enhance the separation of peptides one can also employ two-dimensional or even three-dimensional liquid chromatography, where the separation is based on two or three criteria, respectively.

2.2.3 Peptide ionization in the ion source

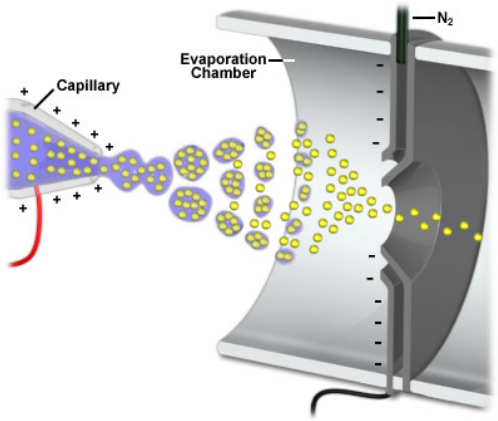


Figure 2.9: Electrospray Ionization Process: The droplets eluating from the needle tip are subjected to an electrostatic field of around 2 kV. This causes the droplets to continuously *explode* into smaller droplets until they only consist anymore of the separate ions that were resolved in them [29].

The two most commonly used techniques used for the ionization of peptides are ESI [16] and MALDI [20]. While MALDI can only be operated offline and is therefore usually used to analyze simple peptide mixtures, ESI is more suited for a shotgun proteomics setup. The reason is that ESI can couple liquid chromatography to mass spectrometry allowing an automated high-throughput analysis.
In ESI a metal needle tip is added at the end of the LC-column, such that the eluate exiting the column passes through it. A voltage of typically 2 kV is applied between the needle and the opposing entrance of the mass spectrometer. The resulting electric field causes the eluate to disperse into a fine spray of charged droplets. With reducing droplet size a point is eventually reached at which the cohesive forces of surface tension are smaller than the repulsive forces between the charges: A so-called *Coulombic explosion* occurs leading to a high number of even smaller droplets [2]. Finally, after repeated iterations of this process, ions of the analytes dissolved in these droplets are produced. The process is illustrated in Figure 2.9.

2.2.4 Tandem Mass Spectrometry

Even after undergoing proteolytic digestion and passing the liquid chromatography column the complexity of most peptide mixtures is still too high to already be sufficiently separated. For this reason an effect called *coelution* occurs, where different peptides elute at the same retention time, which thus enter the mass spectrometer simultaneously. These peptides need to be separated in order to be identified.
To overcome these difficulties tandem mass spectrometry is used, which employs the workflow shown in Figure 2.10. This workflow consists of repeating cycles which in turn consist of the

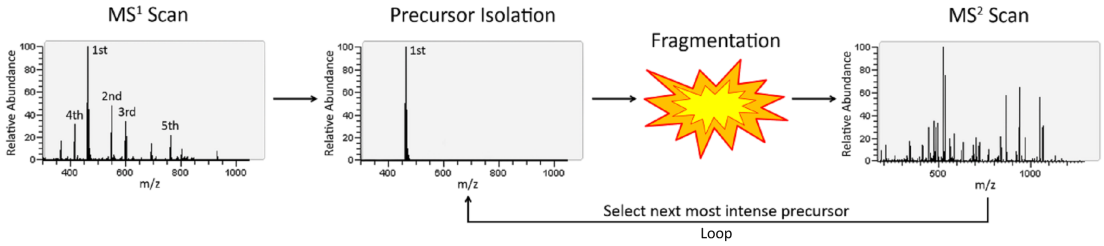


Figure 2.10: The MS/MS-scan cycle: At the beginning of each cycle the most intense precursor ions of a preceding MS1-scan are selected and subsequently isolated one after another for fragmentation (Loop). This gives rise to the MS2-scans of the respective precursor ions showing the resulting fragment ions [44].

following steps:

1. *MS1-scan*:
   The instrument records an MS1-spectrum, i.e. for a defined m/z-range all currently eluting ions (also referred to as the *Total Ion Current* or simply *TIC*) are collected and analyzed. This spectrum shows the measured m/z-values together with the corresponding intensities. The intensity indicates the abundance of ions having the corresponding m/z-value. The higher the resolution chosen in the instrument settings the more well-defined are the resulting peaks (i.e. the smaller their width).

2. *MS2-scan loop - fragmentation of selected precursors*:
   This spectrum is then analyzed and the n most abundant peptides are selected for a subsequent fragmentation analysis. This is also referred to as a *Top-n method*. In order to do so, one by one each of the selected peptides, which are also referred to as *precursors*, is now isolated from the bulk of ions. I.e. a narrow m/z-window of usually 1 to 3 Da with the precursor's m/z-value at its center is applied to the TIC, isolating only ions whose m/z-value lies within this window. After collecting these specific ions for a preset amount of time or until a predefined ion threshold is reached they are sent into the collision cell where they are fragmented. For fragmentation to take place the ions are first activated in the collision cell by elevating their energy levels to an excited state. The natural reduction of the raised energy levels promotes fragmentation of the ions into certain fragment ions (this is explained more detailed later in this section). Finally, the m/z-ratios of these

resulting fragment ions are measured yielding an MS2-scan which shows the abundances for these fragments.

After all n precursor ions have been fragmented and analyzed the cycle repeats with the recording of a new MS1-scan.

Note that the mass analyzer usually also determines the charge state of the precursor ions. This can be done by quickly analyzing isotope patterns found in the spectra. When the charge state z and the m/z-ratio of the ion are known, its molecular mass M can easily be calculated by the following formula:

$$M = (m/z - H^+) \cdot z \tag{2.2}$$

where z its charge state and H^+ the mass of a proton, which is ca. 1.007276 Da.
There are many settings that can be applied to control this general scheme. Besides the above-mentioned possibility of choosing the size of the isolation window, other important parameters include the *dynamic exclusion list* settings. Whenever an ion is chosen for fragmentation its m/z-value is stored on the list for a certain amount of time. None of the m/z-values that are currently listed may be chosen for fragmentation. This way the multiple fragmentation and analysis of peptides that dominate the TIC is avoided, such that there is more time to isolate other less intense peptides. Similarly, one can optionally define an *inclusion list* on which the m/z-ranges are listed that contain peptides of special interest. Only ions whose m/z-value falls into the defined ranges are chosen for fragmentation. This is a useful setting when looking for specific peptides that are part of a complex sample where it is likely that they might be too little intense in comparison to the other ones.

2.2.4.1 Fragmentation of peptide ions by activation leading to MS/MS-spectra

An important step necessary to make tandem mass spectrometry possible is the fragmentation of the peptide ions. Modern mass spectrometers offer different activation types that can be applied to achieve this fragmentation. The different activation types give rise to different kinds of fragment ions, since there are different possibilities to cleave peptides. Depending on the exact cleavage site the following pairs of fragment ions are distinguished:

- a- and x-ions.
- b- and y-ions.
- c- and z-ions.

The a-, b- and c-ions contain the N-terminus, and the x-, y- and z-ions the C-terminus. The fragments are numbered according to the number of amino acids they contain additionally to the N- or C-terminus.
Figure 2.11 shows the possible cleavage sites and gives an illustration of the established nomenclature of the resulting fragment ions [37]. In the following the three most commonly-used activation types are introduced.

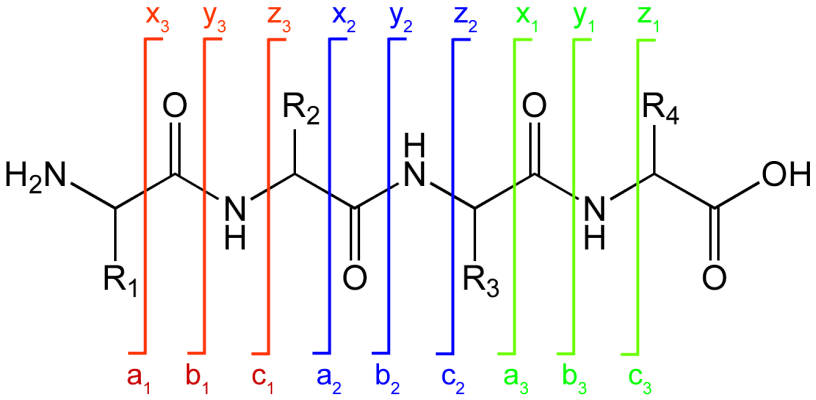


Figure 2.11: Possible Cleavage sites in a peptide and names of the respective fragments [45].

Collision-induced dissociation (CID) CID, sometimes also referred to as collision-activated dissociation (CAD), is the first technique used for the fragmentation of peptides. The precursor ion is kinetically excited which is achieved through collisions with non-reactive gas molecules, such as argon or helium [43]. During each collision the imparted translational energy is converted to vibrational energy that spreads quickly within the molecule via all covalent bonds. As soon as the activation energy required for a certain bond to be cleaved is reached the respective fragment ions are formed. The amide bond chaining the amino acids together is the bond that will most likely break in a peptide ion [50], leading to the above-mentioned b- and y-fragments. Although this technique is very efficient, there are some major drawbacks to be considered: Fragment ions having an m/z-ratio of less than around one third of the precursor ion m/z are not trapped properly due to physical limitations. This effect is commonly referred to as the *low-mass cut off*. Furthermore, the recorded MS2-spectra suffer from relatively low mass accuracy and resolution [31]. The former issue is particularly problematic when applying quantitative proteomics, since the crucial information about the ratios between reporter ions, which have masses lower than 132 Da, is lost.
To overcome the low-mass cut off problem Thermo Fisher Scientific has developed the *Pulsed Q Collision Induced Dissociation* technique (PQD), which is exclusively available in their instruments. This approach promises to yield mass spectra qualitatively comparable to CID spectra without losing the information in the lower m/z regions and thus making them usable for quantitative proteomics.

Higher-energy collisional dissociation (HCD) A more substantial improvement has been achieved with the development of HCD. While on the one hand making the full mass-range of the MS2-spectrum available, this technique additionally provides high mass accuracy and resolution [31]. In principal, it is a higher-energy variant of CID. Precursor ions are also kinetically excited by collision with gas molecules (typically N_2) in order to break up into the characteristic fragment ions.
The resulting spectra have a similar pattern as the ones obtained by CID, i.e. the fragmentation

products that can be measured are the b- and y-fragments. Additionally, however, the fragments from the low mass region, such as the y_1-, y_2-, b_1- and b_2-fragments are also detected and appear in the spectra.

Electron transfer dissociation (ETD) This method is based on the previously developed method of *Electron capture dissociation* (*ECD*) [52]. Like its predecessor ETD has been developed for the use with ion traps, where multiply protonated peptides are confined and subjected to low-energy electrons. The uptake of the electrons is an exothermic reaction and leads to a specific cleavage of the peptides into c- and z-type fragment ions in contrast to the two above-mentioned activation types. However, maintaining the dense accumulation of electrons around the precursor peptide necessary to induce the electron uptake remains technically challenging for some instruments. Therefore, in ETD, contrary to ECD, negative ions (e.g. anthracene) are used in order to achieve the electron uptake of the peptide. Instead of simply capturing a free electron a transfer of an electron from the anion to the peptide takes place, promoted by the fact that anions with very low electron affinities are used [43].
A big advantage of this method over CID or HCD is that the precursor retains labile PTMs as the fragmentation is not achieved by an increase of the internal energy. The downside is the limited applicability to doubly charged precursor ions, which can, however, be overcome by the application of a supplemental activation method to those precursors that are still intact after ETD fragmentation [41].

2.2.5 Analysis of generated mass spectrometry data

2.2.5.1 Spectrum Preprocessing

A typical shotgun-proteomics experiment can generate between 40000 and 50000 MS2-spectra in an average 3h run. This amounts to roughly 3 GB of data. To cope with the vast amounts of data being produced preprocessing steps, such as spectrum-filtering and -manipulation are necessary. All the more since on average only around 50% of the MS2-spectra can be confidently identified.
At this point of the workflow the methods studied and developed in this thesis are settled. They will be explained in chapters 3 and 4.

2.2.5.2 Database Search

In order to automatically identify peptides and proteins, many different database search engines have been developed [7, 9, 15, 35]. The general principle of the search is common to all of them: The protein database is digested *in silico*, i.e. the smaller peptides into which the protein can theoretically be cleaved by a complete cleavage are generated. The experimental MS2-spectra obtained from the mass spectrometer are now compared against the fragment spectra of these theoretical peptides. Depending on the search settings a certain number of possible theoretical peptides are to be considered for the explanation of a given MS2-spectrum. All possible matches are then evaluated using statistical approaches yielding a scoring depending on how probable the matches are. The matching peptides are then ranked according to the score they obtained. This

process is illustrated in Figure 2.12. Notably, this final step of the proteomics workflow is most

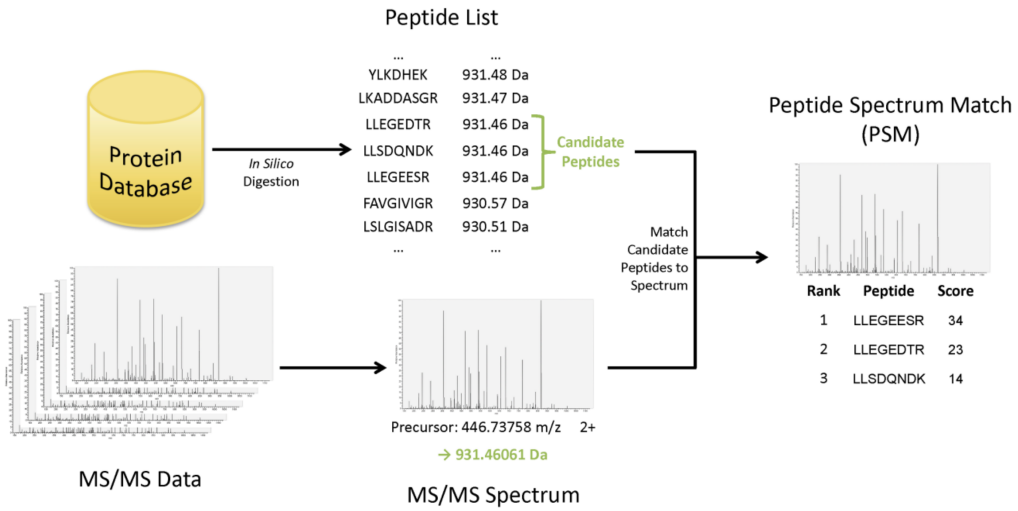


Figure 2.12: Basic workflow of database search engines: The proteins from a selected database are digested in silico into smaller peptides. This yields a list of these peptides and their corresponding molecular masses. For an MS2-spectrum there are a number of candidates whose mass is close enough to the precursor mass, s.t. they are considered as an explanation of the MS2-spectrum. The MS2-spectrum in question is now matched against the theoretical MS2-spectra of the candidate peptides and each match is scored according to how good it is [44].

likely the most versatile one. As it is the crucial step towards making sense of the measured data it is also highly customizable. The first and most important settings are the choice of the database that should be used (human, yeast, etc.), the enzyme that has been used for the proteolytic digestion, the fragment ion types to match and the choice of possible dynamic and static amino acid modifications. These modifications can either be due to chemical derivatisation during sample preparation or post-translational modifications, both leading to a shift in peptide mass. Therefore, they additionally need to be taken into consideration when searching for a matching peptide. Clearly, adding modifications exponentially increases the search space.
Following these basic settings there are - depending on the specific search algorithm - several further parameters that have to be specified. As there are so many different settings and moreover many different settings for different search engines, we only briefly mention the most general ones that are common to most search engines. Most importantly tolerance windows for peptide masses and fragment ion masses can be set. The former narrows the possible candidate peptides to be considered when looking for a match to a certain given mass while the latter narrows the tolerance within which the fragment ions may deviate from the theoretical fragments of a peptide being matched. The tolerance of the search algorithm is usually even further adaptable by setting the maximum amount of missed cleavage sites accommodating for the fact that the proteolytic digestion does not work perfectly.

2.2.5.3 Evaluation of search results

The output of a database search is a list of so-called *peptide spectrum matches* (PSMs) mapping the submitted spectra to peptides. In general, however, we do not have any information about the quality of these results. Elias and Gygy pointed out three reasons why it is necessary to have a means for describing the quality of the search result [14]:

1. The database used for the search may be incomplete i.e. some of the target peptides may not be part of the search space and can never be found.
2. Some spectra may be recorded because of chemical background noise and will therefore lead to false interpretations as random entries in the database might match well enough.
3. In some cases an incorrect peptide matching to a spectrum may be assigned a higher score than the correct interpretation.

It is therefore necessary to establish criteria determining the level of confidence with which the results can be trusted. Moreover, it must be possible to automatically filter search results according to these criteria, such that PSMs that are most probably incorrect can be filtered out. An automatic validation is especially crucial in shotgun proteomics experiments where a vast amount of data is generated and the search results can no longer be validated manually.
A natural quality criterion of database search results is the *false discovery rate* (FDR). In general, the FDR is given by

$$FDR = \frac{FD}{FD + TD} \tag{2.3}$$

where in FD is the number of false discoveries and TD is the number of true discoveries.
In our case the former corresponds to the spectra that have been assigned an incorrect peptide, the latter to the spectra that have been identified correctly. Clearly, $FD + TD$ is equal to the total number of PSMs. However, under normal circumstances it is impossible to determine the FDR of a search result as it is - of course - unknown which of the PSMs are the true and which the false positives, respectively.

To overcome this problem a method of estimating the number of false discoveries in a result list has been proposed: The *target-decoy* search strategy [13, 25]. In this approach two different databases are used for the search. On the one hand there is the original database against which the search is conducted as usual (the *target* database). On the other hand an additional database containing only non-existing peptides is employed (the *decoy* database). Consequently, every PSM mapping a spectrum to a peptide defined in the decoy database is a false positive by definition.
For the estimation of the number decoy hits it is necessary to determine the ratio of decoy to target hits (r_d/r_t) in the search space. This ratio leads to the multiplicative factor f that reflects the decoy bias of the chosen approach, such that, after observing a certain amount of decoy hits (d) and knowing the frequency ratio between target and decoy hits in the database one can calculate the number of false positives [14].

$$f = \frac{1}{r_d} \tag{2.4}$$

$$FP = d \cdot f \tag{2.5}$$

Ideally, the ratio of decoy to target hits is 1:1, such that a random hit may just as likely be a decoy hit as a target hit. I.e. $r_d = r_t = 0.5$ and thus $f = 2$.

In order to achieve this ideal ratio the following criteria must be met by the decoy database with respect to the target database:

1. The amino acid distributions among the peptides should be similar.
2. The protein length distribution should be similar.
3. The total number of proteins should be similar.
4. The number of the theoretical peptides predicted for the digest of the proteins should be similar.
5. The databases may have no predicted peptide in common.

Several approaches to generate the decoy database have been proposed. One possibility is to shuffle the proteins from the target database by randomly rearranging the amino acids they are composed of [19]. Another way would be to create random proteins according to an underlying statistical model to obtain a natural amino acid distribution in the created proteins. However, the most simple approach, reverting the amino acid sequence of the target proteins, proves to yield the best results and is also the most commonly used one in the proteomics community [14].
Using this approach the resulting decoy database clearly fulfills the first four criteria as the respective numbers are exactly the same. Regarding the fifth criterion it cannot be guaranteed that there is absolutely no overlap between the databases. However, the overlap is considerably small. In Figure 2.13 the results of the reversion of the minimally redundant International Protein Index sequence database [21] are illustrated. One can see that almost all of the smaller peptides are found in both databases, but starting from peptides having a length of more than eight amino acids, the overlap is smaller than 0.1 % and thus negligible.

Even when using reversed proteins to generate an almost ideal decoy database, there is still a problem with the target-decoy approach described so far. Performing separate searches in target and decoy databases is more stringent and will, in general, lead to a too high false-positive rate. As in the separate database setting the target and decoy sequences are not competing with each other for the highest-scoring hits, decoy peptides achieve higher scores than they would, if the correct target sequences were part of the search space, as well. This is problematic since these high-scoring decoy hits outscore low-scoring target hits increasing the false-positive rate, although in direct competition the low-scoring target hits might be favored.
The simple solution is to use a concatenated database were both target and decoy sequences are part of the same search space and a sinlge search is run against this database. This dramatically reduces the probability of a decoy hit receiving an increased score while simultaneously reducing the total search time in comparison to the two separate searches.

After analysis of the data via target-decoy search, the decoy hits are removed from the result and the remaining amount of false positives in the target data is then given by:

$$FP_{final} = d(f - 1) \tag{2.6}$$

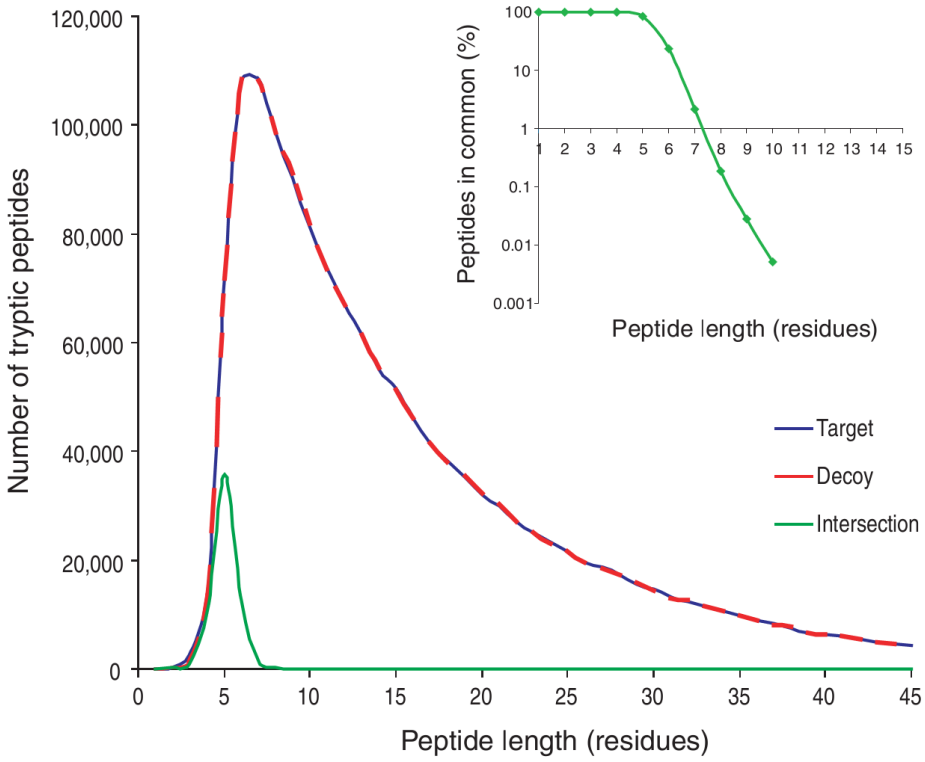


Figure 2.13: Peptide overlap between the human protein sequences within the minimally redundant International Protein Index sequence database (target) and the reversed sequences created out of them (decoy) [13].

Figure 2.14 shows an overview on the target-decoy search workflow (a) and illustrates the distribution of target and decoy hits among all peptide hits for the different ranks assigned by SEQUEST (b-d). The analysis shows that within the first-rank hits the number of reported target sequences is much higher than the number of decoy sequences. In the lower ranks neither sequence set is favored. This confirms the correctness of the approach as the first rank is the most probable hit (the one that is most likely to be correct) while the lower ranks are assigned to less probable sequences and are thus more likely random hits. As intended for random hits, the ratio of target to decoy sequences is 1:1. The aforementioned considerations suggest a fairly simple procedure to obtain a result list filtered to abide a desired FDR. Algorithm 2.1 describes a naive and rough method to set a certain FDR for the results list. The first line is a direct consequence of the fact that for peptides containing less than 8 amino acids there is a considerable overlap in target and decoy sequences. Hence, they violate the precondition that the two databases must have no predicted peptides in common and must therefore be filtered out from the PSM list before setting the FDR. Another problem are hits the search engine has ranked lower than 1. As showed above, for the lower ranks the ratio of target to decoy hits is 1:1 and thus the results are

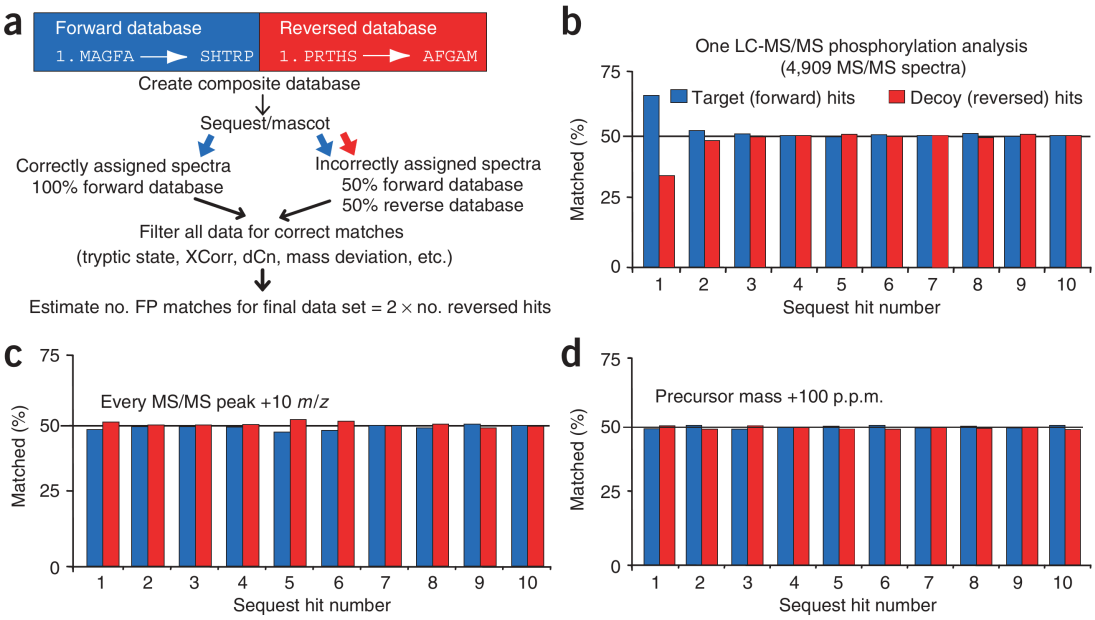


Figure 2.14: a) The target-decoy search workflow. b) Example of PSMs returned by a SEQUEST search. The percentages of matching spectra from both target and decoy database are shown for the top-10 ranking peptide hits. c) The same search but with previously falsified MS/MS spectra. d) The same search but with previously altered precursor masses [4].

mere random hits. Therefore, in order to correctly set the FDR, the lower-ranking results have to be filtered out, as well, which is done in line 2 of the procedure. Afterwards, a score threshold is continuously adapted, until the ratio of target to decoy hits corresponds to the desired FDR. After removal of the decoy hits, the percentage of incorrect hits in the resulting PSM list abides to the set FDR.

Note that it is of course also possible to set other confidence criteria than the search engine score to define the FDR. The following list provides examples of further quality criteria that can be used for defining the confidence of PSMs:

- The ΔScore, i.e. the score difference between the rank 1 and the rank 2 peptides.
- The accuracy of the precursor mass.
- Number of missed cleavage sites.
- How often does the matched peptide match other MS2 spectra.

Moreover, combinations of several criteria are possible. This concept is, for instance, used in the Percolator algorithm [18], which applies a machine learning algorithm to determine the relevant features of the spectra in order to come up with a suitable set of criteria and corresponding weighting factors.

Algorithm 2.1: Set False Discovery Rate

Input : PSMs (list of search results),
FDR (desired false discovery rate)
Output: PSMs, filtered to FDR

```
Remove all hits for peptides smaller than length 8 from PSMs;
Remove all hits for peptides with search engine rank smaller than 1 from PSMs;
FDR_cur ← 1 ;
Score ← 0 ;
while FDR_cur >FDR do
    Score ← Score + Increment ;
    Remove all hits for peptides with a score smaller than Score from PSMs;
    t ← number of target hits ;
    d ← number of decoy hits ;
    FDR_cur = d/t ;
end
Remove all decoy hits from PSMs;
return PSMs
```

Note that regarding identified peptides the FDR is in general larger than the threshold set at the PSM level. This is due to the fact that several correctly identified spectra may be explained by the same peptide. On the other hand, incorrect PSMs distribute randomly over the peptides in the database, as they represent hits by mere coincidence.

2.3 Definition of relevant data structures

In the following a short overview on the data that is output by a mass spectrometer is provided. We list the information that is stored in an output file, and which can generally be seen as given from the viewpoint of the designed algorithms. Unless stated otherwise, the preprocessing algorithms are applied to one MS2-spectrum at a time and thus their input usually consists only of one such spectrum. However, should it be necessary, the rest of the data may be accessed, too, by reading directly from the raw data file.

2.3.1 The .RAW file format

The standard file format of the output files produced by Thermo Fisher Scientific mass spectrometer instruments is the *Thermo Xcalibur file format (.RAW)*. Since this is a proprietary binary format it can only be read by either the Thermo Scientific Xcalibur software or via Thermo's MSFileReader libraries. Parts of the developed programs therefore use the functionality provided by the XRawfile2.dll, version 2.1.1.0.

Most importantly a .RAW file contains all spectra (MS1 and MS2) recorded by the mass spectrometer which can be accessed by their unique running number. The spectra are enu-

merated in the order in which they have been recorded, i.e the first spectrum is the first MS1-spectrum, followed by the associated MS2-spectra, if there are any. Then, the next MS1-spectrum follows, and so on.

Additionally, for each spectrum a collection of additional information, such as retention time, total ion current etc. and certain instrument specific control data and correction factors are stored. Although this information is per se not necessary for the follow-up identification or quantification of peptides, it is still valuable for further analysis to detect flaws in the workflow and devise ways of improving it.

Finally, reports about the defined methods and settings are available for each spectrum.

Note that the crucial information, i.e. the spectrum data itself, is of course also stored in other file formats. The algorithms that are described in this thesis can therefore also be implemented to be operated with these other file formats.
The most popular open data format that should be mentioned is mzML.

2.3.2 Formal definition of mass spectra

For computational means, a spectrum can formally be defined as a list of peaks, where a peak (and the ion indicated by it) is defined by a structure consisting of the following attributes:

Mass Peak data			
Name	Data Type	Abbreviation	Description
Position	double	m/z	The mass per charge ratio m/z of the ion, in a mass spectrum this is where the centroid of the corresponding peak is.
Intensity	double	I	The intensity of the ion current measured in $ions/s$
Charge State	int	z	The charge state of the ion. Note that this information is not necessarily required and the mass spectrometer is not able to determine the charge state for every peak that is recorded!

The column *Abbreviation* denotes the name of the attribute as it will be used in algorithm descriptions within this thesis. E.g. the intensity of a peak p will be referred to as $p.I$.

Instead of having a list of peaks centroids, an alternative, and more precise representation of a mass spectrum can be given by listing not the peak centroid data but the *profile points* each peak consists of. The profile representation requires roughly ten-fold as much memory.

In both cases the peaks (or the profile points) in the list are sorted ascendingly by their m/z values. The following figure shows an excerpt of an actual MS1-spectrum recorded with an LTQ Orbitrap XL instrument. It lists the position and the intensity of one specific peak's profile points.

For MS2-spectra additional information is stored, since an ion from the MS1-spectrum has to be selected to give rise to an MS2-spectrum illustrating the fragments of the respective ion, commonly referred to as the *precursor*. The following attributes constitute the most commonly used precursor data associated with an MS2-spectrum:

```
...
436.79187 391178.8
436.79360 4404735.0
436.79532 11392973.0
436.79704 17311702.0
436.79877 17713638.0
436.80049 12358349.0
436.80222 5574961.0
436.80394 2130796.0
436.80567 1400039.8
...
```

Figure 2.15: Excerpt of an MS1-spectrum schowing the m/z-values and intensities of some profile points.

MS2 Precursor data			
Name	Data Type	Abbreviation	Description
Precursor Position	double	$prec.m/z$	The mass per charge ratio m/z of the precursor ion.
Precursor Intensity	double	$prec.I$	The intensity of the precursor ion.
Precursor Charge State	int	$prec.z$	The charge state of the precursor ion.
Precursor Scan Number	int	$prec.scan$	The number of the spectrum the precursor was selected from.
Precursor Isolation Width	double	$prec.isolation$	The width of the window applied around the precursor m/z determining which m/z-range has been collected for fragmentation.
Activation Type	string	$prec.activation$	Which activation type was used for precursor fragmentation.

Last but not least, there is a collection of header data attached to each spectrum containing further valuable information. The following table provides an overview on the most important parameters that are available in the header.

Spectrum header data			
Name	Data Type	Abbreviation	Description
Scan Number	int	$header.scan$	The running number used to identify the spectrum.
Retention time	double	$header.rt$	The point in time at which the spectrum was recorded
Ion Injection Time	double	$header.ionInject$	The amount of time (usually in ms) the ions have been collected before they were measured.
Low Mass	double	$header.lowMass$	The lower end of the mass range that was considered for the scan.
High Mass	double	$header.highMass$	The upper end of the mass range that was considered for the scan.

CHAPTER 3

MS2 - Spectrum Manipulation

3.1 Motivation

One way to deal with the vast amount of data generated by a shotgun proteomics experiment is the manipulation of MS2-spectra in a preprocessing step preceding the database search. The algorithms described in this section directly target the peak list in a spectrum with the goal of removing unnecessary peaks to facilitate the follow-up analysis done by the search engine. Since the number of possibilities that have to be considered as possible fragment ions is reduced the score of correct identifications is increased. Furthermore, the probability of incorrect identifications is reduced as there are less possibilities for random hits. Of course, these two arguments only hold true, given the assumption that the algorithm works correctly and removes the superfluous peaks it is actually targeting. An additional advantageous side-effect by the reduction of possibilities is the reduction in runtime of the search.

In the following the developed methods for deisotoping and charge-deconvolution of spectra, as well as the ideas they are based on are described. Furthermore, ideas on further possibilities to improve spectrum quality are outlined.

3.2 Deisotoping

As described in section 1.1, a peak in a spectrum (MS1, as well as MS2) indicating the abundance of an ion is usually accompanied by a certain number of shifted peaks. The reason for this effect is the fact that for many chemical elements there are isotope variants that occur naturally. *Isotopes* are atoms whose atomic nuclei contain the same number of protons but a different number of neutrons, therefore having a different atomic mass.
Table 3.1 lists the naturally occurring isotopes for the elements occurring in the proteinogenic[1] amino acids, the first variant of each element is the most abundant (and lightest) one. Note that the mass difference between e.g. a 1 ^{13}C and a 2 ^{13}C isotope is not exactly the mass of the

[1] Proteinogenic amino acids are those amino acids occurring in proteins

Element	Atomic mass (Da)	Natural abundance (%)
Carbon		
^{12}**C**	12.0000000	98.90
^{13}**C**	13.003354838	1.10
Hydrogen		
^{1}**H**	1.007825032	99.985
^{2}**H**	2.014101778	0.015
Nitrogen		
^{14}**N**	14.003074005	99.634
^{15}**N**	15.000108898	0.366
Oxygen		
^{16}**O**	15.994914622	99.762
^{17}**O**	16.999131501	0.038
^{18}**O**	17.999160419	0.200
Sulfur		
^{32}**S**	31.972070690	95.02
^{33}**S**	32.971458497	0.75
^{34}**S**	33.967866831	4.21
^{36}**S**	35.967080880	0.02

Table 3.1: Naturally occurring isotopes of the elements in proteogenic amino acids. The atomic masses are according to [3], the isotopic abundances are taken from [6]

additional neutron[2] the 2 ^{13}C has in comparison to the 1 ^{13}C, but somewhat less. This is due to a phenomenon called *mass defect*, which - simply put - states that the mass of the bound system, i.e. the whole nucleus, is less than the sum of the masses of the unbound system, i.e. the separate protons and neutrons.

Consecutively, two molecules of one and the same compound may have a different total mass as they could contain different isotopes.
We call the variant consisting of the most abundant isotope variatns, i.e containing only light atoms, the *monoisotopic* variant and thus the corresponding peak in the spectra is referred to as the *monoisotopic peak*. As it is the lightest variant of the compound this peak appears the farthest left on the m/z-axis. If the same compound also occurred having one or more of these isotopes exchanged with their heavier variants, this would lead to additional peaks in the spectrum further to the right on the m/z-axis.
Hydrogen and carbon are the most abundant atoms in an amino acid. But as the probability for the occurrence of an ^{2}H atom is only 0.015%, compounds containing them are so rare that they usually do not yield peaks that are intense enough to be distinguished from random noise. On the other hand, with 1.10%, the probability for carbon to occur as the heavier variant (^{13}C instead of ^{12}C) is high enough, s.t. it is likely that among all the C atoms of a certain molecule, there will be one ^{13}C isotope. Likewise, for bigger compounds (containing more C atoms) there

[2]The mass of a neuron is $\sim 1.00866\ Da$

is also a considerable amount of molecules containing two, three or even more ^{13}C isotopes. These heavier molecules also lead to well-visible peaks in the spectrum indicating them.
Note that although for sulfur atoms the probability for a ^{34}S isotope is comparably high, the effect is barely visible in lower m/z-ranges. The reason is that there are only two amino acids containing sulfur, namely cysteine and methionine, which have a natural abundance of 1.9% and 2.3%, respectively [39].

Therefore, as an indication for a certain ion we can usually observe a so called *isotope pattern* in the spectrum, which is composed of peaks indicating variants of the ion containing a different number of ^{13}C isotopes. Since replacing a ^{12}C by a ^{13}C isotope will increase the total mass of the ion by roughly 1 Da, we observe for a monoisotopic peak with an m/z value of v and the charge state c a shifted peak at $v + \frac{n}{c}$ for the ions containing 1 ^{13}C isotopes, a peak at $v + \frac{2n}{c}$ for the ions containing 2 ^{13}C isotopes, etc. Here, the n stands for the mass difference between ^{12}C and ^{13}C which is ca. 1.003354838. Figure 3.1a shows an isolated example of a typical isotope pattern, consisting of the monoisotopic, as well as the 1 ^{13}C and the 2 ^{13}C peaks, as it could occur in a spectrum.

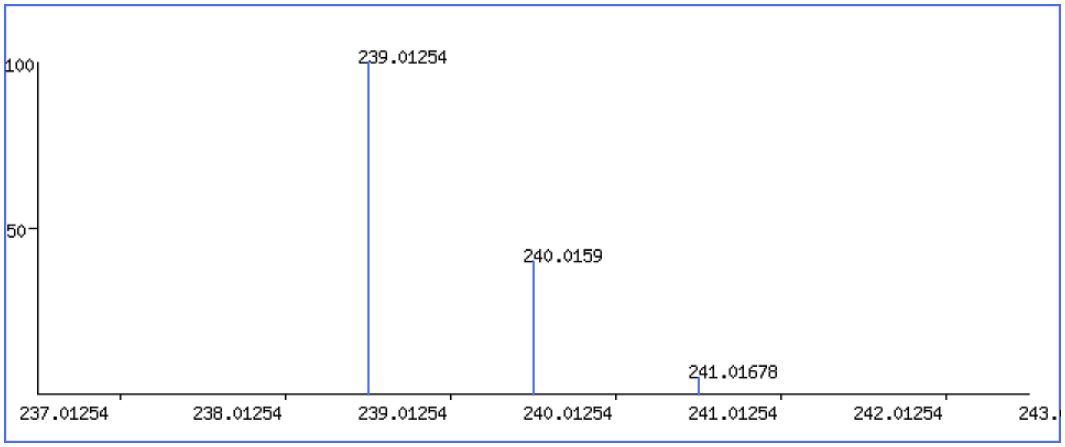
As common search engines do not take additional isotope peaks into account but only rely on the monoisotopic peak to identify fragment ions, it is a reasonable approach to see these superfluous peaks as noise and remove them from the spectrum.

We therefore experimented with several ideas, starting from the approach by Savitski et al. [38], in order to develop a reliable way of determining isotope patterns in MS2-spectra and cleaning them thereof, s.t. only the monoisotopic peaks remain. Every peak, that is not accompanied by an isotope pattern, should remain untouched. This general procedure is referred to as *Deisotoping*. Figures 3.1a and 3.1b show an isotope pattern before and after deisotoping.
The following sections give a detailed description of the ideas that we experimented with along with a short summary of the results achieved with the respective method.

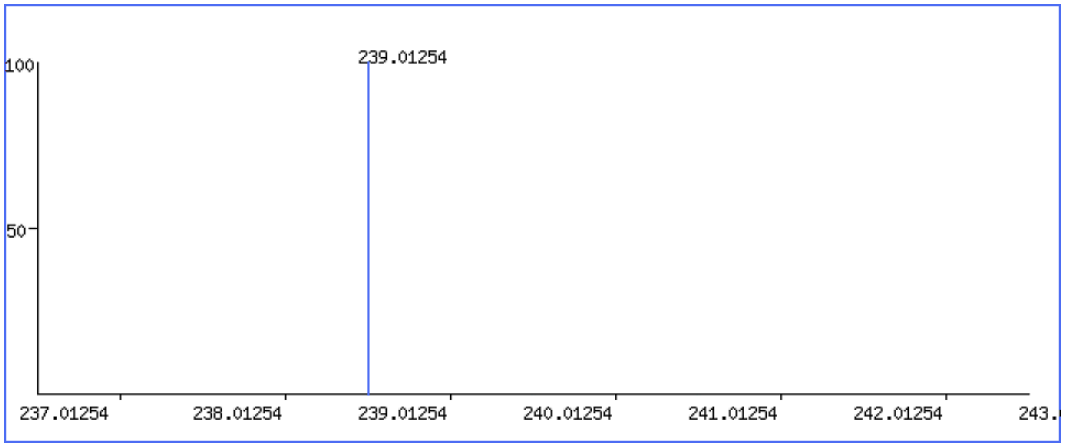
3.2.1 A simple Deisotoping algorithm

The deisotoping approach we started with has originally been implemented by its authors as part of their H-Score approach [38], which we do not further explain here, as it is out of the scope of this thesis. As the aim of H-Score was the introduction of a new scoring scheme for high mass accuracy data, the deisotoping method incorporated in it is a suitable starting point for the development of new sophisticated deisotoping algorithms for the high mass accuracy spectra we want to optimize.

The algorithm makes use of the above-mentioned distance of $\frac{n}{c}$ between consecutive isotopes for a c-fold charged analyte ion. It traverses the peak list backwards, i.e. starting from the peak with the highest m/z-value towards the peak with the lowest one. For each peak p_i it inspects, if there is a previous peak p_j having the distance $\frac{n}{c}$ for $c \in [1..p]$ (with p being the precursor charge). Of course, inaccuracies in the data have to be accounted for, which is why a certain tolerance value has to be applied, the authors used $\pm 0.025\ Da$, here. Additionally, a simple plausibility criterion is checked: From the two peaks in question, the intensity of the first peak, i.e. the one with the greater m/z-value, may be at most twice as high as the intensity of the second one. If these two criteria are fulfilled, the second peak is marked for removal, which is done by simply setting its intensity to 0. In the end, every peak with intensity 0 is removed



(a) Example of a typical isotope pattern. The monoisotopic peak of an ion at $239.01254\ Da$ (also referred to as the ^{12}C or the 0 ^{13}C peak) is followed by the 1 ^{13}C peak at $240.0159\ Da$ and the 2 ^{13}C peak at $241.01678\ Da$



(b) The same situation after deisotoping. Only the monoisotopic peak remains in the spectrum.

Figure 3.1: Simple isotope pattern, before (a) and after (b) deisotoping

from the peak list.
Algorithm 3.1 gives a pseudo-code of the described procedure.

This algorithm is simple and straightforward. Indeed, a preprocessing of the MS2-spectra using this method improves the identification rate compared to the search of unprocessed spectra (for a detailed comparison see section 5.2). We could even see an improvement for CID spectra which have lower mass accuracy. Obviously, the algorithm leaves room for several improvements having some considerable drawbacks:

1. The tolerance window of $\pm 0.025\ Da$ is too generous considering the high resolution of HCD spectra.

2. The algorithm marks detected peaks too quickly, i.e. whenever two peaks fit the criteria, the second peak is marked. This procedure can be problematic when overlapping patterns

Algorithm 3.1: Deisotoping as implemented by Savitski et al. [38]

Input : A list of $(m/z, \mathsf{I})$-pairs, peaks, of length N ascendingly ordered by m/z
Structure containing precursor ion information, prec
A value defining the accuracy tolerance of the isotope search, tolerance
Output: Deisotoped peaks list

```
1  n ← 1.0033548;                    // mass difference between 12C and 13C
2  for i ← N to 1 do
3      for j ← i − 1 to 1 do
4          for c ← prec.z to 1 do
5              if Abs(peaks[i].m/z − peaks[j].m/z - n) ≤ tolerance and
               peaks[i].I > peaks[j].I/2 then
6                  peaks[j].I ← 0 ;              // mark this peak for removal
7              end
8          end
9      end
10 end
11 foreach peak p in peaks do
12     if p.I = 0 then
13         Remove p from peaks
14     end
15 end
```

occur. A simple example is depicted in Figure 3.2.

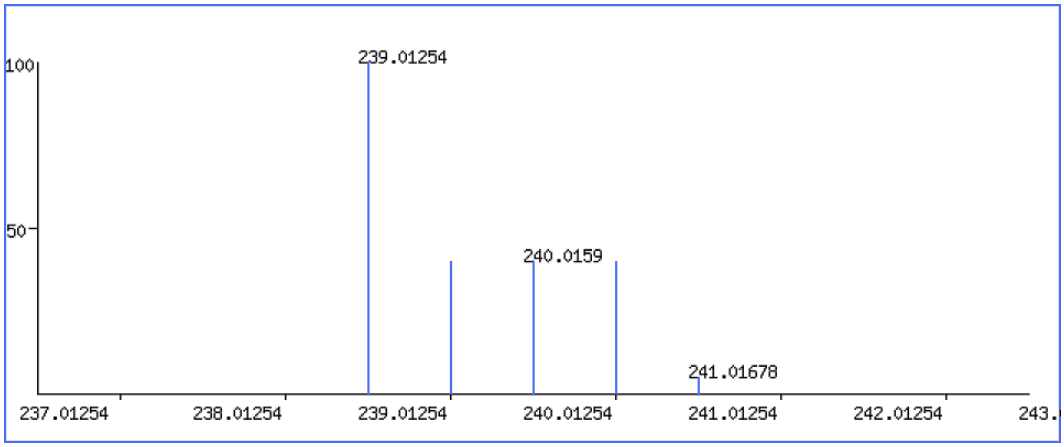


Figure 3.2: Overlapping isotope patterns: Here, the three annotated peaks form an isotope pattern for charge state $z = 1$. The algorithm, however would also remove the two additional peaks, since they would fit in an isotope pattern with charge state $z = 2$.

3. The first peak falling in the tolerance window is marked as an isotope peak. However, there may be another peak having a distance closer to the expected one. Of course, this peak would then be a better match and should be the only one that is marked. Instead, every peak that could be a possible isotope is marked.

4. Two consecutive peaks having the correct distance are, in general, seen as isotopes. The chosen plausibility criterion, only rejecting such two peaks, if the second peak is at least twice as high as the first one, is a too simple way of modeling natural isotope abundances. Thus, almost every time when two peaks have the correct distance, they are recognized as indicating isotopes and consequently, the second peak is removed.

Generally spoken, these issues can be summarized by the observation that the algorithm is somewhat inaccurate and too little restrictive. It simply removes too many peaks from a spectrum.

We therefore try to overcome these issues by modifying the original algorithm accordingly.

3.2.2 Improving the original algorithm

The first issue can be solved rather easily by simply reducing the size of the tolerance window. To this end we implemented a variable user-definable tolerance window allowing for more restrictive tolerances. However, test runs with a tolerance of $\pm 0.02\ Da$, $\pm 0.015\ Da$, $\pm 0.01\ Da$, $\pm 0.005\ Da$ and $\pm 0.001\ Da$ did only show a slight but not significant improvement in the number of PSMs. As can be seen in the experimental results in section 5.2, the optimal deisotoping tolerance seems to be around 0.015 Da, or 15 mmu, for HCD spectra.

In order to cope with the second problem, we slightly modify the existing algorithm. Instead of marking peaks for removal, as soon as they are found, the possible isotope patterns for the current peak under consideration of every charge state are stored in a list. Additionally, each pattern is evaluated by a simple scheme, resulting in a score for this pattern. After performing the search for all possible charge states, the pattern with the best score is then chosen and its peaks (except for the monoisotopic peak, of course) are marked for removal.
The score of a possible isotope pattern P consisting of n isotope peaks beside the monoisotopic one is given by the sum of the distance of each peak p_i's observed position $p_i.pos_o$ from its expected theoretical position $p_i.pos_t$.

$$score(P) = \sum_{i=1}^{n} |p_i.pos_o - p_i.pos_t| \tag{3.1}$$

This method assumes the best-scoring possible isotope pattern to be the actual one. It is therefore more cautious than the original approach as it keeps peaks in the spectrum that are less likely part of the sought isotope pattern.

In a similar way the third issue of the original algorithm could be handled. When looking for possible isotope peaks within the tolerance window, we do not accept the first match but rather look further for any other possible matches. Again, all possible peaks that would match this specific isotope peak are stored in a list. Finally, the best-fitting peak is chosen, i.e. the peak whose deviation from the expected distance to the previous isotope peak is smallest.

However, this approach only reduced the number of PSMs and was therefore discarded.

3.2.2.1 Expected isotope-ratio determination by averagine modeling

The last issue is the most difficult one to deal with, since modeling the isotope abundance ratios requires an adaptive approach that is strongly dependent on the number of C-atoms contained in the respective molecule. Naturally, we do not have any information about the composition of the molecule prior to the search. However, its mass is known, which allows an estimation of the ratios.

In order to have an admissible estimation of the isotope ratios of a given molecule, the theoretical isotope intensity ratios of a poly-averagine [39] (i.e. a peptide consisting of multiple averagines) of the same mass are used. *Averagine* is the average amino acid, a theoretical construction that has been developed based on the statistical occurrences of the amino acids from the PIR protein database [49]. This model amino acid has the molecular formula $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$ and an average mass of 111.1254 Da. As described above, the more C atoms there are in a peptide, the higher the probability that one, two, three or more of these C atoms are actually ^{13}C and not ^{12}C isotopes.

This is easy to see. We can formulate the question for the number of ^{13}C isotopes in a molecule as a serial experiment in which the C-atoms are independently examined one by one yielding either ^{13}C or ^{12}C as result. This is exactly what is modeled by the Binomial distribution. Thus, the probability for having k ^{13}C-isotopes in a molecule containing n C-atoms in total is given by

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{, where } p = 0.011 \tag{3.2}$$

Statistics suggests that these are also the average intensity ratios we can see when studying the respective isotope peaks. Figure 3.3 illustrates the probabilities for peptides consisting of an increasing number of averagines to contain 0, 1, 2, or 3 ^{13}C isotopes. The Figure suggests that a more fine-grained acceptance criterion is needed. Certainly, the criterion chosen in the original algorithm is fulfilled by any two isotopes of molecules up to 3333,762 Da (the mass of a molecule consisting of 30 averagines). However, the problem lies in the opposite cases, when this criterion seems far too relaxed. Consider for example the ^{12}C and the 1 ^{13}C isotopes of a peptide consisting of two averagines. The ratio between the intensities of the two peaks should be around 1 to 10. Now, if we have a spectrum in which there are two peaks at roughly 220 Da having the distance expected from two isotope peaks but a ratio of 1 to 1, which considerably differs from the expected one, the original algorithm would still consider these peaks isotopes and delete the second of those from the spectrum.

These considerations suggest that an algorithm that uses a more sophisticated acceptance criterion taking the actually expected ratios into account should be more accurate and thus yield better results.

We therefore analyzed 103431 MS2-spectra recorded by a Thermo QExactive instrument measuring a 1 μg HeLa sample to find out how well this theoretical model fits to actual data. The analysis was done in the following way:

1. For every isotope pattern found in the spectra select the monoisotope peak and calculate its uncharged mass m.

2. Calculate how many C-atoms there are in a poly-averagine of the same mass.

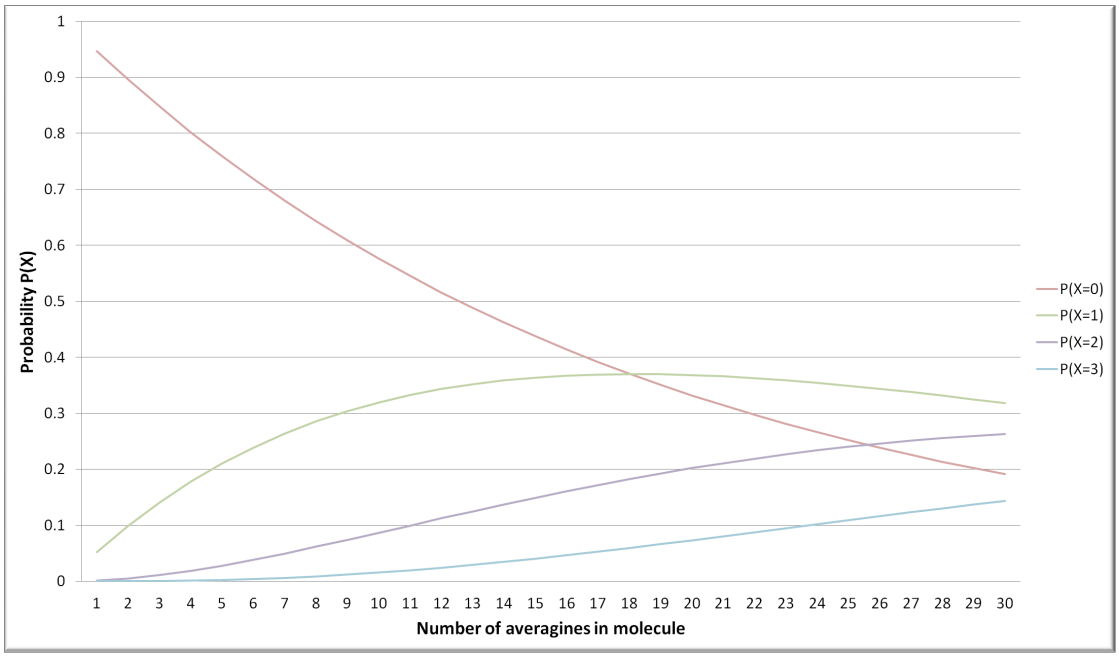


Figure 3.3: Probability distribution for the number of ^{13}C-isotopes in a poly-averagine molecule depending on the number of averagines in the molecule.

3. Calculate the theoretical intensities p^0_{theo}, p^1_{theo}, p^2_{theo} and p^3_{theo} for the 0 ^{13}C, 1 ^{13}C, 2 ^{13}C and 3 ^{13}C isotopes, respectively using formula 3.2. Note that the calculated number n of C-atoms is usually not an integer. Therefore we perform a linear interpolation between the rounded down and rounded up values, i.e.:

$$P(X = n) = P(X = \lfloor n \rfloor)) \cdot (n - \lfloor n \rfloor)) + P(X = \lceil n \rceil)) \cdot (1 - (n - \lfloor n \rfloor))), \quad (3.3)$$

where $\lfloor n \rfloor$ and $\lceil n \rceil$ denote the applications of the floor and the ceiling functions to n, respectively.

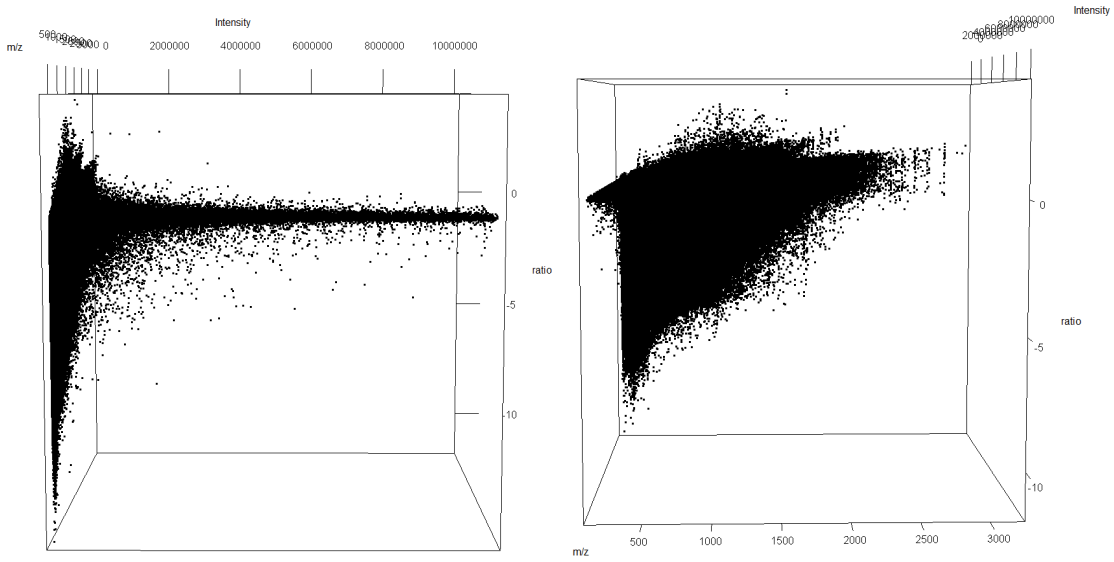
4. Calculate the sum of the first 4 peaks in the pattern in order to determine their percentages of this total intensity: p^0_{obs}, p^1_{obs}, p^2_{obs} and p^3_{obs}.

5. Compute the ratio $r_i = \frac{p^i_{theo}}{p^i_{obs}}$ for $0 \leq i \leq 3$.

Clearly, in an optimal fit every such calculated ratio is exactly 1. To obtain an extensive overview the data was plotted in the three dimensions, m/z and intensity (both taken from the monoisotope peak), as well as the ratio. Note that the ratio has been log2 - transformed for the plots, s.t. the range is compressed and optimally-fitting data points lie on the plane ratio $= 0$. The following plots show the data collected for each of the four isotopes separately, focusing on the intensity-ratio plane (left) and the m/z-ratio plane (right). The study of these plots leads to two conclusions:

1. The ratio is clearly correlated to the intensity and is better, the higher the intensity is.

2. The ratio is not directly correlated to the m/z-value of the peaks. Although the plots might suggest some kind of correlation, the different trends for the certain isotope peaks show that this is indeed just a reflection of the correlation with the intensity. The explanation is simple. As Figure 3.3 also shows, for ions with lower masses, the intensity of the 1 ^{13}C peak - and moreover the 2 ^{13}C and 3 ^{13}C peaks - is relatively small in comparison to the ^{12}C. Especially if the isotope peaks are so small, that they are close or beneath the noise level, it is therefore very likely that the percentage of the total intensity is heavily distorted.
   Thus, a decent fit can only be seen for the ^{12}C and the 1 ^{13}C isotopes. Regarding the other two, a general tendency can be seen but the variance is too high.

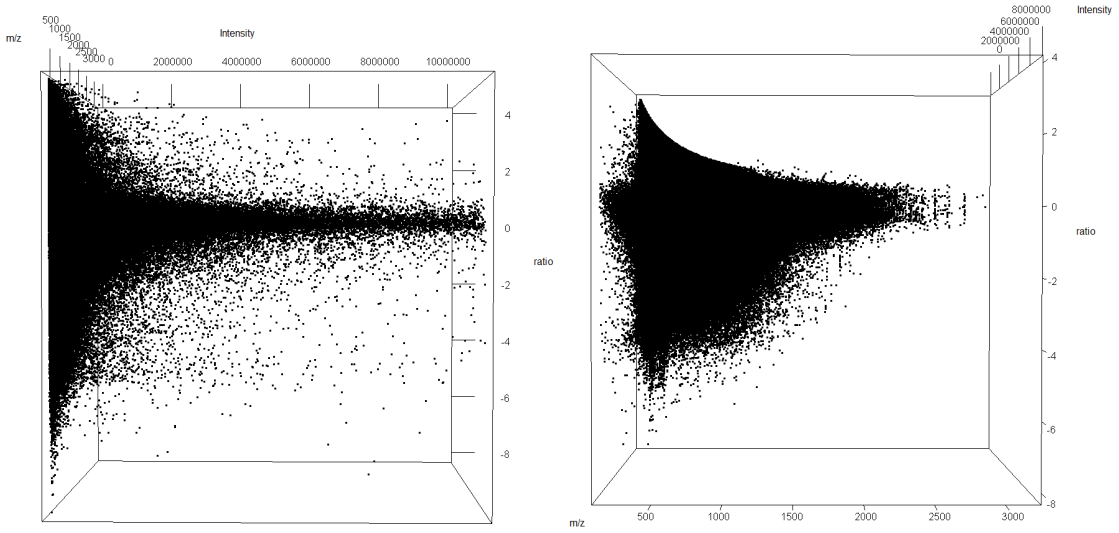


(a) (b)

Figure 3.4: 3d-plot showing the ratios of the measured against the theoretical intensities for the ^{12}C isotopes.

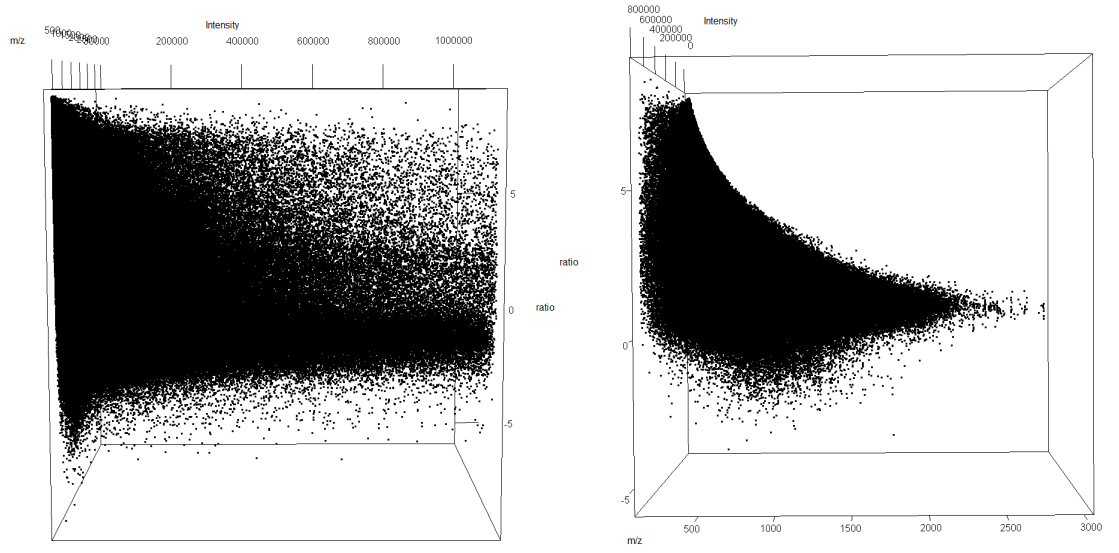
Consequently, we developed an algorithm that makes use of this correlation with the aim of rejecting patterns for which the isotope ratios are not good enough. The principle is the same as for the conducted spectra analysis. Additionally, after step 5 as described above, the computed ratios r_i are used for evaluation. However, as a conclusion of the analysis, we only consider ratios r_0 an r_1, as only for those the correlation is good enough to be used as a quality criterion: We demand of both ratios to be greater than a given threshold value. If this is the case, the isotope pattern is accepted, and all but the monoisotope are removed.

Unfortunately, testing several ratio thresholds, as well as intensity thresholds did not significantly improve the performance in comparison to the original approach. Moreover, it seems



(a) (b)

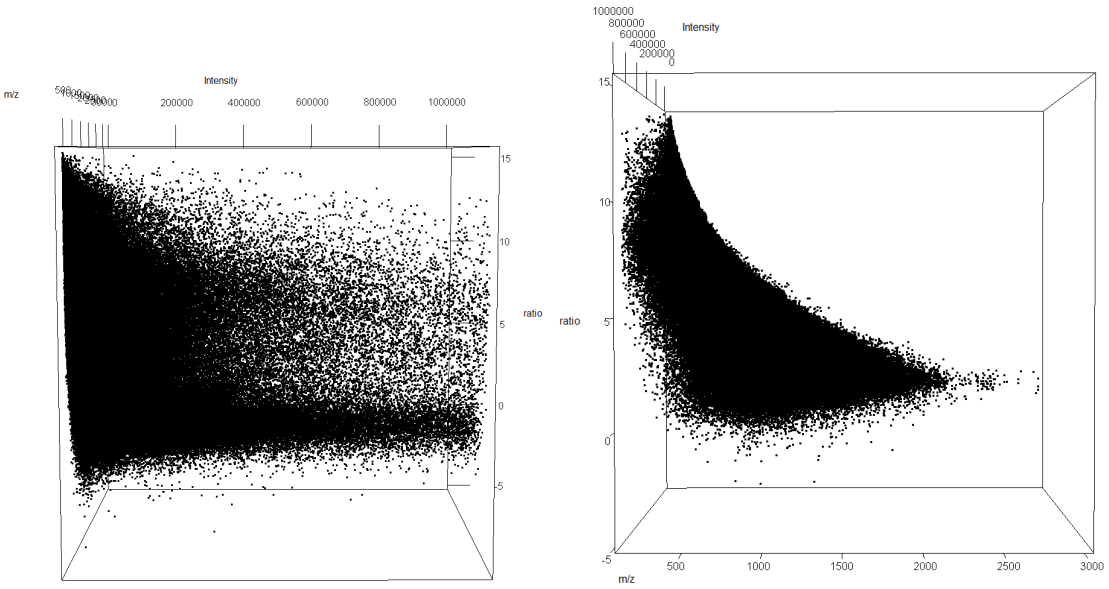
Figure 3.5: 3d-plot showing the ratios of the measured against the theoretical intensities for the 1 ^{13}C isotopes.



(a) (b)

Figure 3.6: 3d-plot showing the ratios of the measured against the theoretical intensities for the 2 ^{13}C isotopes.

sample-dependent, if the method yields any benefit at all or if it even reduces the number of



(a) (b)

Figure 3.7: 3d-plot showing the ratios of the measured against the theoretical intensities for the 3 ^{13}C isotopes.

PSMs (see section 5.2). For CID-spectra the application of the averagine approach visibly reduces the results.

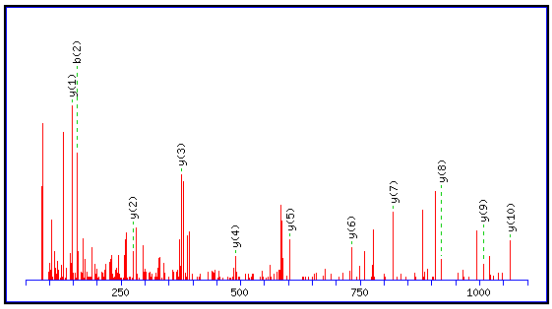
3.2.2.2 Reasons for the superiority of the simple approach

Unfortunately, none of the attempts to improve the original deisotoping algorithm was really successful. Despite the fact that the number of peaks incorrectly classified as isotopes was decreased the overall result did not improve significantly. Even worse, as shown in section 5.2, the number of PSMs has even decreased in some cases compared to the simpler algorithm.
Naturally, the question arises why this is the case.

The most thorough way to answer this question is the comparison of the same spectrum both after applying the original approach and a modified one. Moreover, this needs to be done for spectra for which the search results differ. This way, significant differences can be determined. An examination of Mascot search results quickly led to the conclusion that the original algorithm works better towards the way the search engine evaluates the PSMs. In principle, if a peak is matched, the removal of any peaks in its vicinity will increase the score. As the score is based on probabilities, the removal of *competing* peaks decreases the probability of the matching peak being a random match. Since usually only a small part of the peaks are matched, it is much less likely to accidentally remove a peak that will be matched than a random non-important one. Therefore, on average, the achieved scores are better with the original algorithm than with a more restrictive one. Figures 3.8, 3.9 and 3.10 illustrate three examples of a comparison between the

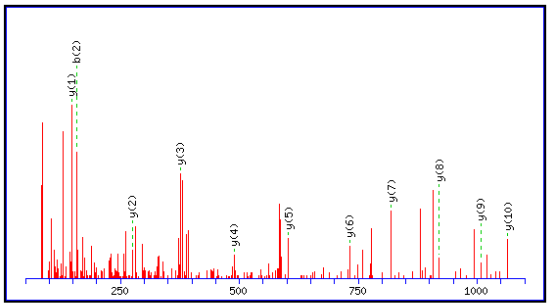
original algorithm run at 25 mmu and 15 mmu isotope deviation tolerance. Figure 3.8 shows a spectrum where the sole fact that more peaks have been removed improved the score. Figures 3.9 and 3.10 show excerpts of spectra where additionally the higher tolerance led to the matching of another fragment ion, which is not found in the more restrictively processed spectra.

To prove the principle that the general removal of peaks can have a higher benefit than



Monoisotopic mass of neutral peptide Mr(calc): 1162.58298
Fixed modifications: Carbamidomethyl (C)
Ions Score: 41 **Expect:** 0.0023
Matches (Bold Red): 11/40 fragment ions using 51 most intense peaks

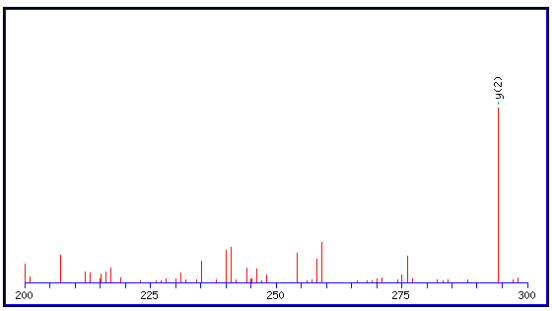
(a) 15 mmu



Monoisotopic mass of neutral peptide Mr(calc): 1162.58298
Fixed modifications: Carbamidomethyl (C)
Ions Score: 51 **Expect:** 0.00027
Matches (Bold Red): 11/40 fragment ions using 41 most intense peaks

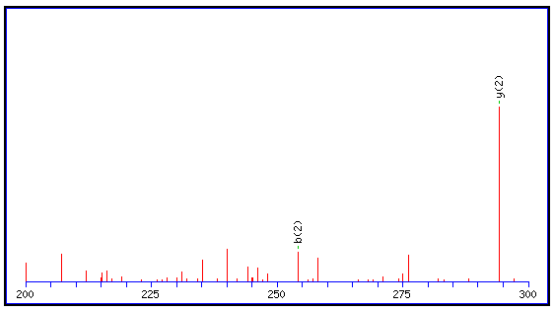
(b) 25 mmu

Figure 3.8: Comparison of Mascot search results: The 25 mmu variant of algorithm leaves the spectrum much less dense than the 15 mmu one. Especially the lower m/z-range is much sparser.



Monoisotopic mass of neutral peptide Mr(calc): 1578.69838
Fixed modifications: Carbamidomethyl (C)
Ions Score: 40 **Expect:** 0.004
Matches (Bold Red): 13/48 fragment ions using 17 most intense peaks

(a) 15 mmu

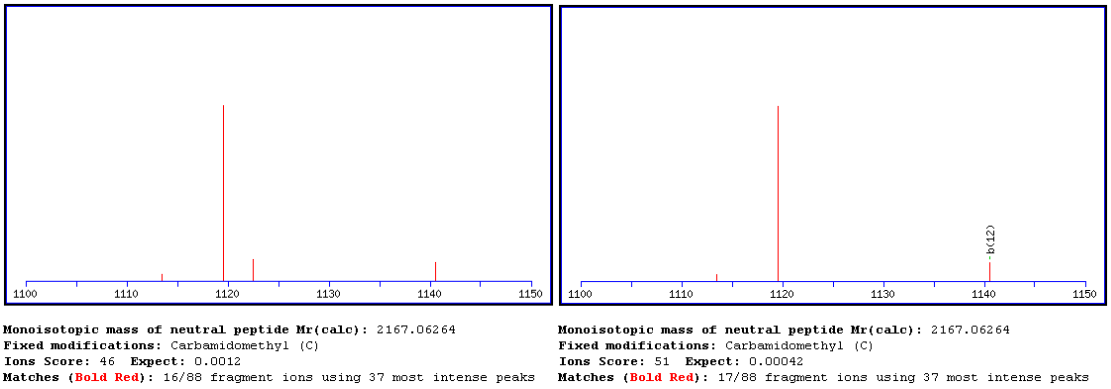


Monoisotopic mass of neutral peptide Mr(calc): 1578.69838
Fixed modifications: Carbamidomethyl (C)
Ions Score: 50 **Expect:** 0.00039
Matches (Bold Red): 18/48 fragment ions using 46 most intense peaks

(b) 25 mmu

Figure 3.9: Comparison of Mascot search results (zoomed in to mass range 200 to 300 Da): With 25 mmu Mascot was able to match the b_2 ion due to the removal of the slightly higher peak at around 259 Da. In the 15 mmu variant the peak is not removed and obstructs the matching of the b_2 ion.

being more careful with which peaks to remove we experimented with even greater tolerance



(a) 15 mmu (b) 25 mmu

Figure 3.10: Comparison of Mascot search results (zoomed in to mass range 1100 to 1150 Da): With 25 mmu Mascot was able to match the b_{12} ion due to the removal of the slightly higher peak at around 1122.5 Da. In the 15 mmu variant the peak is not removed and obstructs the matching of the b_{12} ion.

windows up to 100 mmu. Interestingly we do not observe a considerable performance drop, which confirms the stated assumption.
The evidence is even clearer in the case of lower resolution CID spectra. Although the studied methods are targeting HCD (high-resolution, high accuracy) data and we did not expect the more accurate deisotoping approaches to perform very well on CID spectra, it is nevertheless remarkable that here, the results are exactly the other way around: The sloppier the method the better it performed. Moreover, the sophisticated averagine approach failed completely and was in one case outperformed even by the unprocessed spectra.

We therefore conclude that there is not too much gain in peptide identification performance to be obtained from MS2-deisotoping procedures. Moreover, as the more sophisticated approach does not always yield better results, we recommend to employ the original algorithm as this one does not increase the runtime unnecessarily.

3.3 Charge-Deconvolution of spectra

Another typical phenomenon that can be observed is the occurrence of peaks in different m/z-ranges indicating the same fragment ion. This is due to the different charge states that an ion can have after fragmentation, depending on where on the original peptide the protonations have taken place [23]. As the m/z-scale shows higher-charged ions in the lower ranges and lower-charged ions in the upper ranges this leads to overcrowded spectra containing much more information than necessary. The m/z-value of an ion is computed the following way:

$$m/z = \frac{M + zH^+}{z} \tag{3.4}$$

Algorithm 3.2: Charge-Deconvolution as implemented by Savitski et al. [38]

Input : A list of (pos, I)-pairs, peaks, of length N ascendingly ordered by pos
Structure containing precursor ion information, precursor
Output: Charge-deconvolved peaks list

```
1 p ← 1.007276;                                  // molecular mass of a proton
2 for i ← 1 to N do
3 |   if peaks[i].z > 1 then
4 |   |   peaks [i].pos ← peaks [i].pos ·(peaks[i].z − 1)H+ ;
5 |   |   peaks [i].z = 1 ;
6 |   end
7 end
```

where M is the molecular mass of the ion, z its charge state and H^+ the mass of a proton, which is ca. 1.007276 Da.
E.g., if a fragment having the molecular mass of 1500 Da would occur singly-, doubly- and triply-charged, the respective spectra would contain peaks at 1500 Da, 751 Da and 501 Da, respectively.

Most search engines do not consider higher charged ions (Mascot, for example, only considers singly- and doubly-charged fragment ions) as this would exponentially increase the search space. Therefore it seems a promising approach to remove multiply-charged replicas of the same fragment ion from the spectra and only keep the singly-charged ones.

Clearly, for charge-deconvolution to be possible, the charge states of the peaks have to be known. It is therefore recommended to perform the Deisotoping procedure beforehand, if the peaks in the spectra are not annotated with charge states or it is unclear if they are.

3.3.1 A simple Charge-Deconvolution algorithm

For the charge-deconvolution we followed the same approach as for the deisotoping. We started from the existing algorithm developed for H-Score [38].
The principle of the algorithm is quite simple. For the charge-deconvolution only one simple iteration through the peak-list is necessary. For each peak with a charge state greater than 1 the singly-charged m/z-value is then calculated according to eq. 3.4:

$$m_{sc} = m/z \cdot (z-1)H^+ \tag{3.5}$$

The peak is then simply replaced by a new peak at position m_{sc} having the same intensity but the charge state $z = 1$. Algorithm 3.2 illustrates this procedure. Unfortunately, the preprocessing with this method only yields a negligible improvement in the identification performance (see section 5.2). Here, two possible problems play a role:

1. The "conversion" of a peak indicating a multiply-charged ion into a peak indicating a singly-charged peak does not take into consideration, if there already is a peak for the

singly-charged ion. As a possibly already existing peak will in general not have exactly the same m/z-value as the shifted one, this will lead to peaks referring to the same ion which are, from the search engine's point of view, competing with each other. This will however, lower the resulting score for matched peaks in the same m/z-range.

2. Even though the charge-deconvolution thins out the lower m/z-regions where the peaks of the multiply-charged ions can be found the higher regions are populated with additional peaks, which leads to new competing peaks there. The obtained benefits from the thinning might therefore just be canceled out by the suffered score decrease through the new competing peaks.

Although the second problem clearly cannot be eliminated we experimented how the algorithm performs, if the search engine only considers singly-charged ions.
Here, we do observe a greater increase in PSMs, however the improvement can still not be considered significant and, in particular, it is lower than having the search engine make use of the doubly-charged ions.

Furthermore, we tried to reduce the negative effect of newly introduced peaks by consolidating an already existing peak with the shifted one, i.e. by solving the first problem. To achieve this we changed the algorithm to look for an already existing peak at the position ($\pm$ 0.01 Da) the converted peak would be shifted to. If such a peak exists the two peaks' intensities are compared. Whichever has the higher intensity remains in the spectrum, the other one is removed. Additionally, the remaining peak's intensity is increased by the intensity of the removed peak. This revised method is given in algorithm 3.3. In fact, this change had no effect whatsoever on the resulting PSMs. The only measurable change was very small and only occurred for one of our three test instances (see 5.2) that it can be considered negligible.

3.4 Results and Future Work

As a result of the analyses described in this section we decided to implement the simple existing deisotoping and charge-deconvolution procedures in a plugin for the Proteome Discoverer software, version 1.3.0.339, by Thermo Fisher Scientific: The *MS2-Spectrum Processor* node, that can be integrated in the workflow. As there was no significant performance increase by the more sophisticated methods, we chose the simple approaches to not necessarily waste computation time. The software allows the setting of a desired tolerance value for the deisotoping method, as well as the general choice, if deisotoping and/or charge-deconvolution should be applied to the MS2-spectra.
For testing purposes and possible analyses of spectra it is also possible to switch to the averagine modeling approach. However, as we have showed in this section, this will not necessarily yield a higher PSM count and is therefore not recommended as a standard procedure in the processing workflow. Similarly, we implemented the modified charge-deconvolution method for testing purposes, as we believe it is more exact than the original one.

As the more sophisticated algorithms did not yield significant improvements in comparison to the simple algorithm we assume that the design of even more complex methods would probably also not be very promising. The final algorithms implemented in the plugin are therefore

Algorithm 3.3: Modified Charge-Deconvolution algorithm

Input : A list of (pos, I)-pairs, peaks, of length N ascendingly ordered by pos
Structure containing precursor ion information, precursor

Output: Charge-deconvolved peaks list

```
1  p ← 1.007276;                                   // molecular mass of a proton
2  for i ← 1 to N do
3      if peaks[i].z > 1 then
4          M_singly ← peaks[i].pos ·(peaks[i].z − 1)H+ ;
5          singlyChargedPeak_cur ← FindClosestPeak (M_singly) ;
6          if Abs (singlyChargedPeak_cur - M_singly) ≤ 0.01 then
7              if peaks[i].I > singlyChargedPeak_cur .I then
8                  peaks[i].pos = M_singly ;
9                  peaks[i].I = peaks[i].I + singlyChargedPeak_cur .I ;
10                 peaks[i].z = 1 ;
11                 singlyChargedPeak_cur .I = 0 ;
12             end
13             else
14                 singlyChargedPeak_cur .I = peaks[i].I + singlyChargedPeak_cur .I ;
15                 peaks[i].I = 0 ;
16             end
17         end
18     end
19     else
20         peaks[i].pos ← peaks[i].pos · (peaks[i].z − 1)H+ ;
21         peaks[i].z = 1 ;
22     end
23 end
24 foreach peak p in peaks do
25     if p.I = 0 then
26         Remove p from peaks
27     end
28 end
```

simple and not very time consuming, such that they can be generally applied in the workflow without having to consider excess runtime or memory limitations.

Naturally, there are many more spectrum-manipulation approaches one could apply, yet, as we demonstrated in this section, the main benefit with current search engines lies in the removal of peaks. To this end, another possibility would be the removal of neutral loss[3] peaks

[3]The term *neutral loss* describes the loss of an uncharged compound like water or ammonia, which is reflected in the spectrum by another peak of specific m/z-offset from the original fragment ion.

of “confirmed” ions, i.e. ions for which an isotope pattern has been found. Furthermore, we believe that a noise filtering method that clears the low-intensity areas of the spectrum would also increase the number of PSMs.
Probably another promising approach is the recalibration of MS2-spectra. For this approach a two-searches strategy would have to be implemented where the first search generates a result list that serves as an input to recalibrate MS2-spectra using high-confidence identifications. The recalibrated spectra can then be submitted to the second, actual, search. Section 4.4 describes this idea with respect to precursor recalibration. However, for this approach to work a search engine would be needed that rewards accurate fragment ions.

CHAPTER 4

Precursor Precision Improvement

4.1 Motivation

In general, mass spectrometers have to take decisions on which precursor to pick for fragmentation analysis within a few milliseconds. Moreover there is no spare computation time left to look back at the chosen precursors and try to get some additional information about them. Therefore, the typical approach is to take the data that is available at the moment of the precursor-picking. However, using not only the information from the MS1-spectrum the precursor originated from but also from other spectra in which the precursor can be found should allow for a more exact calculation of the precursor mass.
The reason is simple. As we are dealing with measured values which are subject to errors we can model the set of all measurements of a single precursor mass by a normal distribution whose mean is the correct value of the mass. And clearly, the obtained result of averaging multiple measurements will, in general, be closer to the correct value than simply considering just one single measurement. This is even more true for spectra obtained by HCD. As described in section 2.2.4.1 these spectra feature high mass accuracy and resolution.

We thus have developed a method that can be applied as an intermediate processing step after the acquisition of the spectra and before the spectra are submitted to the database search engine. This way, the necessary data can be gathered from the complete raw file, i.e. all (MS1-) spectra that have been recorded can be incorporated in the analysis, without being subject to real-time constraints the instrument software has to deal with when choosing precursor ions for fragmentation. The algorithm processes one MS2-spectrum after another with the basic input being - besides the possibility to access all raw data generated from the respective sample - the m/z-value of the precursor ion associated with the spectrum. After running the algorithm a new, more reliable m/z-value is the expected output and replaces the value originally associated with the MS2-spectrum.

4.2 Method

4.2.1 Basic Principle: 3d-Peaks

The key idea behind the developed approach is the 3d-peak. This concept has already been successfully applied in the MaxQuant software and it has been shown that the number of peptide identifications could be increased dramatically [7]. However, the proposed algorithm does not always extract 3d-peaks correctly and tends to be too restrictive. Furthermore, we employ different methods of improving the precision of the resulting m/z-value, as a part of the mass improvement employed in MaxQuant is based on the detection of SILAC pairs or triplets [1], which cannot be applied in general, i.e. for non-labeled samples. Last but not least, the algorithm in MaxQuant has been developed and tested for CID spectra, for HCD spectra different approaches might be necessary.
Nonetheless, the general concept of 3d-peaks promised to serve as a solid basis for the development of our algorithm.

As described above the aim is to collect further precursor data from different MS1-spectra than the one the precursor ion was selected from. The logical step to take is therefore to study precursor peaks not only in the m/z-intensity plane but to include the retention time as the third dimension and consider three-dimensional peaks.
The basic method for 3-dimensional peak-picking is simple:

1. Search the precursor in the original MS1-spectrum and gather its profile points.

2. Starting from this spectrum, traverse backwards through the MS1-spectra and search for the current precursor in each of them, until the precursor cannot be found anymore.

3. Starting from this spectrum, traverse forwards through the MS1-spectra and search for the current precursor in each of them, until the precursor cannot be found anymore.

$\Rightarrow$ The 3d-peak is constructed from the separate 2d-peaks formed by the profile points that were obtained from the respective MS1-spectra.
In the following, we will refer to the peak formed by the profile points as a 2d-peak, while the "peak" obtained by regarding the sequence of the 2d-peaks is referred to as the 3d-peak. Figure 4.1 illustrates the connection between 2d- and 3d-peaks.

Although the basic idea behind the algorithm sounds simple, especially the construction of the 3d-peak and the decision where the peak ends is a complex problem, which will become clear in sections of this chapter.
The details of the algorithm are described in the following.

4.2.1.1 2d-Peak Detection

The first basic element of the algorithm is the detection of a 2d-peak in a single MS1-spectrum. Recall that a 2d-peak is defined by the profile points of which there are around 10,000 to 20,000

[1] SILAC is a method of quantitative proteomics, in which peptides are *labeled* with light and heavy forms of a certain amino acid. Based on this labeling the measured peptides can then be quantified.

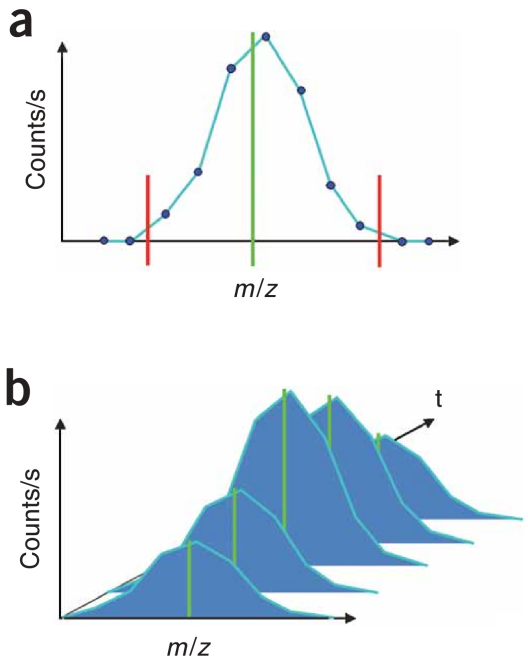


Figure 4.1: *a)* A 2d-peak formed by the profile points in a specific MS1-spectrum. The green line depicts the centroid of the peak, the two red lines define its ends. *b)* A 3d-peak composed of five 2d-peaks that were found in different MS1-scans adjacent to each other [7].

in an average MS1-spectrum, depending on the resolution. The challenge is therefore to collect only the profile points that are in fact part of the same peak. As the peak shapes are usually Gaussian, the peaks can be separated very confidently from each other leading to a very simple method to collect exactly the profile points that form one peak. Starting from the profile point closest to the expected centroid value, the list of profile points is traversed both towards lower and higher m/z regions until a point is reached whose intensity value is 0 or a local intensity minimum was detected, indicating a so-called *split peak*. Note that the former criterion can be applied since on the edges of the peaks there are usually several profile points with intensity 0 included in the list. This is due to the baseline- and noise-filtering applied by the instrument software while writing the data to the raw file. If this were not the case, and only peak data were stored in the list, the two neighboring peaks could still be separated via the second criterion. An example of where a split peak would be detected can be seen in Figure 4.2. Naturally, when searching for the point closest to the given centroid, a tolerance window around this centroid has to be applied, s.t. cases where there is no point inside the window result in no matching peak. A good choice is a window width of 5 to 10 ppm.
It was proposed that this simple approach of detecting the 2d-peak is sufficient for spectra recorded by high precision instruments [7].

After the profile points of a peak have been collected, its centroid is calculated. As the peak shape is Gaussian it is an accurate approach to fit a Gaussian function to the collected profile

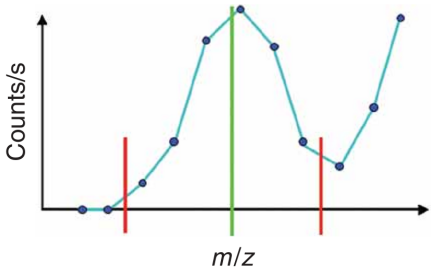


Figure 4.2: Example of a 2d-split peak: Note that the margin between the two bordering peaks can easily be found by searching for the local minimum [7].

points and to determine the centroid of this function. In their supplemental material Cox et al. [7] suggest the following formula to calculate an approximation of the centroid m a Gaussian function fitted to the profile points would have

$$m = \frac{1}{2}\frac{(L_0 - L_1)m_{-1}^2 + (L_1 - L_{-1})m_0^2 + (L_{-1} - L_0)m_1^2}{(L_0 - L_1)m_{-1} + (L_1 - L_{-1})m_0 + (L_{-1} - L_0)m_1} \tag{4.1}$$

where m_{-1},m_0, m_1 are the m/z-values (i.e. the positions) of the three central profile points, and L_{-1}, L_0, L_1 are the natural logarithms of the corresponding intensities. For the rare cases that a peak is defined by less than two profile points the centroid is given by the average of their m/z-values weighted by their intensities.

A closer analysis of this formula confirms that it can indeed be used to determine the centroid of a Gaussian. The general Gaussian function is defined by

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}} \tag{4.2}$$

where $a, b, c > 0$.
Applying the natural logarithm to this function yields a quadratic function

$$lnf(x) = ln\, a - \frac{1}{2c^2}x^2 + \frac{b}{c^2}x - \frac{b^2}{2c^2} \tag{4.3}$$

which in its general form is given by

$$lnf(x) = ax^2 + bx + c \tag{4.4}$$

Now, in our case the domain of this parabola is m/z and the function range are the logarithmized intensities. We can determine the three unknown coefficients, a, b and c of the function by solving a system of three linear equations using the given three m/z-values m_{-1}, m_0, m_1, and the corresponding logarithmized intensities L_{-1}, L_0, L_1.

$$L_{-1} = a\, m_{-1}^2 + b\, m_{-1} + c \tag{4.5}$$
$$L_0 = a\, m_0^2 + b\, m_0 + c \tag{4.6}$$
$$L_1 = a\, m_1^2 + b\, m_1 + c \tag{4.7}$$

Note that the centroid of a parabola is equal to its maximum, which can easily be obtained by differentiation, i.e. m can be found by solving

$$m = -\frac{b}{2a} \tag{4.8}$$

The constant part of the function c, is therefore irrelevant and we can proceed in the following way:
By subtracting 4.7 from 4.6, 4.5 from 4.7 and 4.6 from 4.5 we obtain

$$\begin{aligned} L_0 - L_1 &= a(m_0^2 - m_1^2) + b(m_0 - m_1) && (4.9)\\ L_1 - L_{-1} &= a(m_1^2 - m_{-1}^2) + b(m_1 - m_{-1}) && (4.10)\\ L_{-1} - L_0 &= a(m_{-1}^2 - m_0^2) + b(m_{-1} - m_0) && (4.11) \end{aligned}$$

Now, we divide 4.9 by $(m_0 - m_1)$, 4.10 by $(m_1 - m_{-1})$ and 4.10 by $(m_{-1} - m_0)$ which yields

$$\begin{aligned} \frac{L_0 - L_1}{(m_0 - m_1)} &= a(m_0 + m_1) + b && (4.12)\\ \frac{L_1 - L_{-1}}{(m_1 - m_{-1})} &= a(m_1 + m_{-1}) + b && (4.13)\\ \frac{L_{-1} - L_0}{(m_{-1} - m_0)} &= a(m_{-1} + m_0) + b && (4.14) \end{aligned}$$

In order to isolate a, we subtract 4.13 from 4.12:

$$\begin{aligned} & \frac{L_0 - L_1}{(m_0 - m_1)} - \frac{L_1 - L_{-1}}{(m_1 - m_{-1})} = a(m_0 - m_{-1}) \\ a &= \frac{L_0 - L_1}{(m_0 - m_1)\cdot(m_0 - m_{-1})} - \frac{L_1 - L_{-1}}{(m_1 - m_{-1})\cdot(m_0 - m_{-1})} \\ &= \frac{(L_0 - L_1)\cdot(m_1 - m_{-1}) - (L_1 - L_{-1})\cdot(m_0 - m_1)}{(m_0 - m_1)\cdot(m_1 - m_{-1})\cdot(m_0 - m_{-1})} \\ &= \frac{(L_1 - L_0)m_{-1} + (L_{-1} - L_1)m_0 + (L_0 - L_{-1})m_1}{(m_0 - m_1)\cdot(m_1 - m_{-1})\cdot(m_0 - m_{-1})} && (4.15) \end{aligned}$$

Now, we use equation 4.14 in order to isolate b

$$b = \frac{L_{-1} - L_0}{(m_{-1} - m_0)} - a(m_{-1} + m_0) \tag{4.16}$$

and insert a as calculated above:

$$\begin{aligned} b &= \frac{L_{-1} - L_0}{(m_{-1} - m_0)} - \frac{(m_{-1} + m_0)\cdot((L_1 - L_0)m_{-1} + (L_{-1} - L_1)m_0 + (L_0 - L_{-1})m_1)}{(m_0 - m_1)\cdot(m_1 - m_{-1})\cdot(m_0 - m_{-1})} \\ &= \frac{(L_{-1} - L_0)\cdot(m_0 - m_1)\cdot(m_1 - m_{-1})\cdot(m_0 - m_{-1})}{(m_{-1} - m_0)\cdot(m_0 - m_1)\cdot(m_1 - m_{-1})\cdot(m_0 - m_{-1})} \\ &\quad - \frac{(m_{-1}^2 - m_0^2)\cdot((L_1 - L_0)m_{-1} + (L_{-1} - L_1)m_0 + (L_0 - L_{-1})m_1)}{(m_{-1} - m_0)\cdot(m_0 - m_1)\cdot(m_1 - m_{-1})\cdot(m_0 - m_{-1})} \end{aligned}$$

$$
\begin{aligned}
= & \frac{(L_{-1}-L_0)\cdot(m_{-1}-m_1)\cdot(m_0-m_1)}{(m_0-m_1)\cdot(m_1-m_{-1})\cdot(m_0-m_{-1})} \\
& -\frac{(m_{-1}+m_0)\cdot((L_1-L_0)m_{-1}+(L_{-1}-L_1)m_0+(L_0-L_{-1})m_1)}{(m_0-m_1)\cdot(m_1-m_{-1})\cdot(m_0-m_{-1})} \\
= & \frac{(L_0-L_1)m_{-1}^2+(L_1-L_{-1})m_0^2+(L_{-1}-L_0)m_1^2}{(m_0-m_1)\cdot(m_1-m_{-1})\cdot(m_0-m_{-1})}
\end{aligned} \tag{4.17}
$$

Finally, a and b can be inserted into the derivative, 4.8:

$$
\frac{1}{2}\frac{(L_0-L_1)m_{-1}^2+(L_1-L_{-1})m_0^2+(L_{-1}-L_0)m_1^2}{(L_0-L_1)m_{-1}+(L_1-L_{-1})m_0+(L_{-1}-L_0)m_1} \tag{4.18}
$$

which is exactly the proposed formula.

We conclude that the proposed formula indeed yields the centroid of a fitted Gaussian to the three profile points.

The intensity of a 2d-peak is given by the area under the Gaussian curve. However, because the intervals between the separate profile points are nearly constant, the *total intensity*, i.e. sum of the separate intensities is proportional to the area under the Gaussian [7]. For further computations the centroid-m/z and the total intensity are stored for the each 2d-peak.

It is therefore possible to obtain reasonably good approximations of the centroid and the intensity of a Gaussian peak without actually performing a runtime-expensive linear regression of a Gaussian function to the profile points.

Note that this is also done for the original precursor m/z received as input, as we want to make sure that the centroid, that is being used further on by the algorithm is from that 2d-peak only. As the instruments apply a proprietary peak-picking algorithm it is not clear how the original centroid value has been calculated.

4.2.1.2 3d-Peak Centroid

After calculating the centroid for each 2d-peak the actually sought 3d-peak centroid is calculated. This is done by simply computing the average of the n 2d-peak centroids ($m_i, 1 \leq i \leq n$) weighted by the corresponding total intensities ($I_i, 1 \leq i \leq n$):

$$
m/z_{3d} = \frac{\sum_{i=1}^{n} m_i I_i}{\sum_{i=1}^{n} I_i} \tag{4.19}
$$

This is the value that is output by the algorithm replacing the original precursor m/z-value.

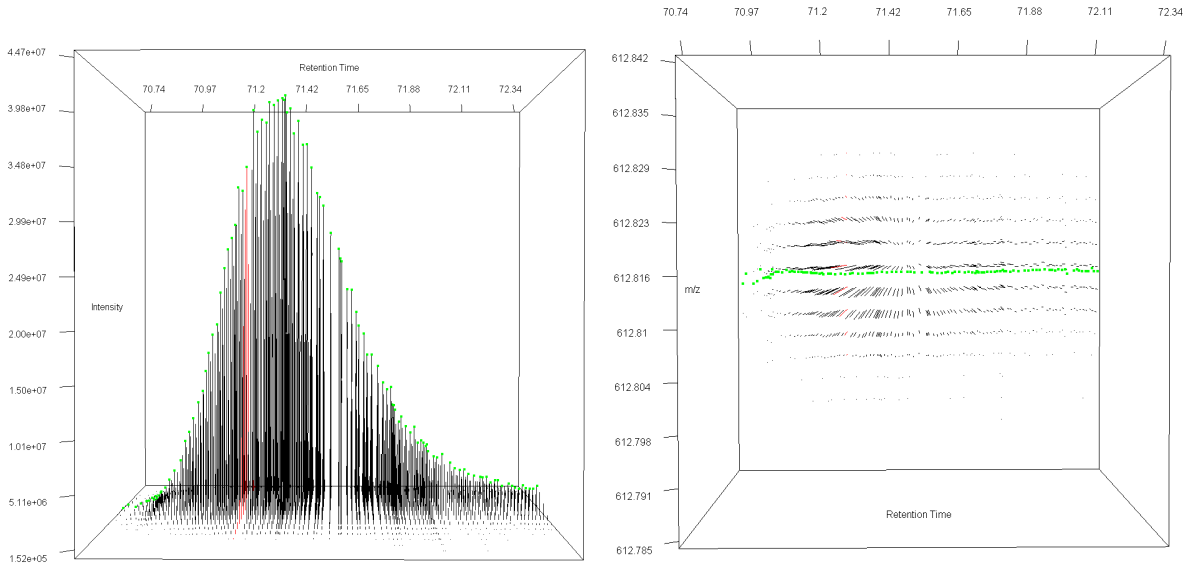
The idea behind this calculation is the fact that the measured values become less accurate the closer the intensity is to the (technical) noise level. Therefore, the higher confidence in m/z-values stemming from highly intense peaks is expressed by the greater weight.

4.3 3d-Peak Reconstruction

A crucial aspect of the algorithm has not been discussed yet: *How is a 3d-peak constructed?* Since the retention time dimension of the 3d-peak should reflect the elution behavior of the

corresponding peptide, it is necessary to reconstruct the elution profile in order to obtain a proper 3d-peak. The 3d-peak's profile in the retention time-intensity plane has an almost Gaussian shape, as well. However, the rising edge of the curve is slightly steeper, whereas the falling edge is much flatter compared to a regular Gaussian curve. Additionally, the falling edge of the quasi-Gaussian part of the peak usually passes over to a more or less distinctive tailing which - in extreme cases - can be several times as long as the main part of the peak. The reason for this phenomenon lies in the gradient used in the HPLC (see section 2.2). When the point is reached where the concentration of the organic compound is high enough to loosen the ions of a specific peptides from the solid phase they begin moving towards the end of the column where they will elute. If at this point the concentration were kept at a constant level, the eluting ions would result in a Gaussian-shaped elution profile. However, as the concentration keeps rising the rising hydrophobicity forces the ions even faster towards the end of the column. Furthermore, with a continuously lowering probability to remain bound to the stationary phase the elution time becomes prolonged. This results in the described deviations from a Gaussian profile.
Regarding the retention time-m/z plane, a 3d-peak is ideally Gaussian-shaped, which is obvious as all the 2d-peaks of which it consists are Gaussian-shaped. Figure 4.3 shows an example of an extracted peak featuring a typical elution profile as described above, as well as almost no deviations along the m/z-axis. In this and the further example plots the following data is visualized:
For each example a 3-dimensional plot in the dimensions *Retention Time (RT) in min*, *Intensity* and m/z was made of the respective extracted 3d-peak. The figures consist of two parts, one showing the RT-Intensity plane, the other one the RT-m/z plane. Each series of black lines symbolizes the profile points of the 2d-peak that was extracted from the spectrum recorded at the respective retention time. The green dots on top of the 2d-peaks mark the centroid m/z-value calculated for the peak. Finally, the red lines indicate the profile points taken from the original precursor spectrum. Note that in general the shape of a 3d-peak is by far not that perfect as theory expects it to be, as the measured intensities and m/z-values are subject to distortion. As described above, the m/z-values become more noisy the closer to the noise level they are. Unfortunately, the intensity values themselves are also subject to distortion that increases with decreasing intensity. At a closer look one can also see these deviations in the RT-Intensity plot in Figure 4.3, where even peaks in the apex area of the elution profile are deviating visibly from the expected ideal shape. Especially low-abundant peptides leading to low-intensity peaks are therefore subject to heavy distortion, i.e. inaccuracy in both measured dimensions, which complicates an effective peak-reconstruction. This demands for an algorithm capable of detecting split peaks without being to restrictive as smaller deviations are usually within the normal range.

Reconsider the above-mentioned procedure. The starting point is the spectrum in which the precursor has been found. Now, in the spectra recorded previously and later we continue the search for the precursor and add the extracted 2d-peaks, until the precursor cannot be found anymore. However, it still needs to be clarified what it means that a precursor is found or not found in a certain spectrum. There is indeed no easy answer to this question as in complex samples a myriad of peptides elute simultaneously and the phenomenon of overlapping peaks is not rare. In order to enable the algorithm to take well-founded decisions the information about



(a) RT-Intensity plane

(b) RT-m/z plane

Figure 4.3: Example of an ideal 3d-peak, having an approximately Gaussian-shaped elution profile. In the end of the peptide's elution time, the peak has a distinctive tailing. Regarding the RT-m/z plane the observed values follow a straight line.

the 3d-peak shape that has been reconstructed so far, needs to be available. For this reason the centroid and intensity values of each 2d-peak are calculated right after it has been extracted from a spectrum. Based on these values and the values of the previously added 2d-peaks, the algorithm can now decide if this new 2d-peak is most probably part of the 3d-peak and should thus be added or if it should be rejected.

The relevant parameters that need to be set for the peak detection are listed and briefly explained in table 4.1. How exactly each of these parameters is applied in the algorithm is explained subsequently in this section and the following subsections. Initially, it needs to be determined, if the spectrum currently searched contains a suitable *candidate peak*, i.e. a 2d-peak which is close enough to apply the further plausibility checks.
The spectrum is searched by applying the detection algorithm described above using as m/z-value around which the search is performed the centroid of the peak found in the previous spectrum (or the original precursor value, for the first scan). If a profile point is found within the tolerance window the resulting 2d-peak is the candidate peak to be attached to the sequence of 2d-peaks forming the current 3d-peak the detection. Otherwise, no candidate peak is considered for this spectrum.

If a candidate peak has been extracted, the following criteria are checked to determine, if it was given rise by the same peptide and should thus be attached to the current 3d-peak:

Name	Data Type	Abbreviation	Description
Tolerance	double	tol	Defines the maximal deviation from the original centroid. I.e. a tolerance window of $\pm tol$ is applied around the original centroid.
Minimum Profile Points	int	n_{min}	A candidate peak needs at least n_{min} profile points in order to be considered.
Skip Scans	int	$skip$	The amount of spectra that can be skipped in case the candidate peak does not match the criteria.
Points for Regression	int	n_{reg}	The amount of 2d-peaks that are taken into account for the linear regression used in peak-profile tracking.
Points for Mass Drift Detection	int	n_{drift}	The amount of 2d-peaks that are taken into account for the detection of slow mass drifts.

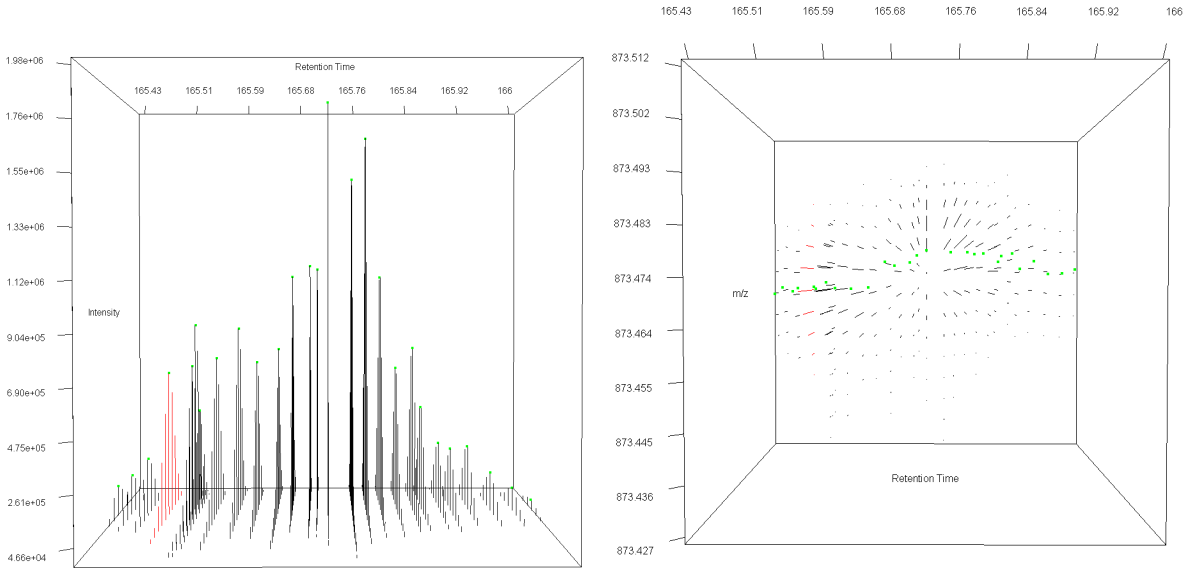
Table 4.1: 3d-peak reconstruction: Relevant settings

4.3.1 Minimum Profile Points in 2d-Peak

If the extracted candidate peak does not consist of more than the set minimum amount of profile points it is rejected. The reason is that due to noise- and baseline-correction performed by the mass spectrometer the extracted peak represents only the apex region of the actual peak. Consequently, the m/z-values of the remaining profile points, and moreover the centroid, will most probably not be very accurate.
One might argue that the part such a very low intense peak will contribute to the weighted-average centroid of the 3d-peak is insignificant, which is certainly true. This check is nevertheless valuable as every increase in precision - no matter how small - can be an advantage. Additionally, this check rather reduces than increases the runtime of the algorithm. When the peak reconstruction reaches the margins of the 3d-peak, the 2d-peaks clearly become less intense. This means that as soon as a 2d-peak is encountered, that is so small that it no longer consists of a high enough amount of profile points, the 3d-peak is already ending. It is therefore reasonable to stop the search and not to continue until no 2d-peak at all can be found anymore. This is especially true in the tailing area of the 3d-peak, where there can still be plenty of spectra containing a low intense 2d-peak that is, strictly seen, part of the same 3d-peak. Added up these peaks' intensities could have a considerable amount of weight to observably alter the overall centroid.
In the light of these considerations this criterion becomes much more valuable as it appears at a first glance.

4.3.2 Peak to Peak m/z-Deviation

Another simple criterion that is checked is the deviation in the m/z-dimension. If the candidate peak's centroid deviates from the one of the previous peak by more than the set tolerance, the peak is rejected.
The aim behind this check is to detect sudden shifts in the m/z-dimension of greater magnitude. As for modern instruments the measured m/z-value of peptides remains relatively stable, such a deviation clearly indicates that the candidate peak does not indicate the same peptide as the ones collected so far in the 3d-peak. It is therefore safe to assume that the current 3d-peak has ended and a different one is just beginning. Figure 4.4 shows an example of an extracted peak containing such a sudden m/z-shift. In this case a closer look at the elution profile also indicates that at around 165.65 min on the RT axis actually another 3d-peak starts while the original 3d-peak is already ending.



(a) RT-Intensity plane

(b) RT-m/z plane

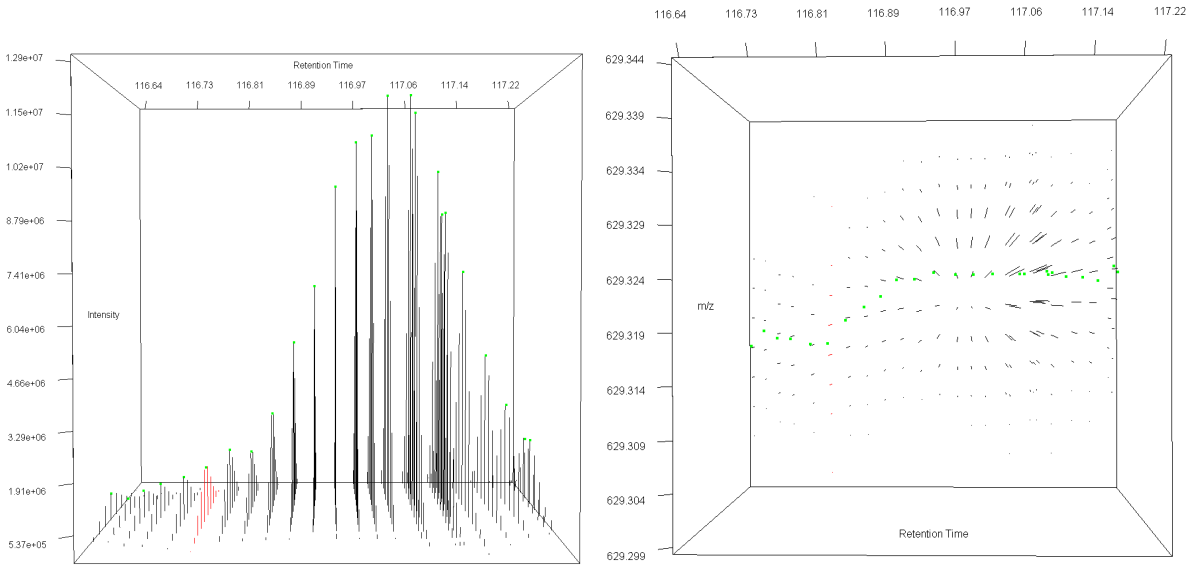
Figure 4.4: Example of a 3d-peak, in which a sudden m/z-shift occurs around 165.65 min on the RT-axis

4.3.3 Continuous m/z-Deviation

Unfortunately, it is not always that easy to detect overlapping peaks. In many situations there is indeed a true co-elution of peptides meaning that the resulting 3d-peaks will overlap heavily. The generation of one single 3d-peak from two co-eluting peptides would lead to a deformed peak shape in the m/z-axis, whereas the elution profile of the peak does not necessarily reflect the overlap. The reason is that the respective intensities for the two peptides add up leading to a elution profile that represents, in fact, the sum of the eluting ions of both peptides. In situations

where the overlap is big, one does not even observe abrupt mass shifts, as a big overlap means that the profile points of both peptides overlap in many spectra leading to many *joint* 2d-peaks. Therefore the respective centroids of these peaks only gradually drift from the m/z-range of one peptide, to the other one. Figure 4.5 shows an example of a 3d-peak whose m/z axis contains a continuous drift upwards. One could assume by inspecting the elution profile that there are two peaks, the first one ending at around 116.9 min, the second one starting at around 116.75 min. Thus, 5 to 6 2d-peaks are overlapping, which is reflected in the m/z-drift, which ranges over the said 5 to 6 2d-peaks.

For the sake of detecting such drifts the algorithm analyzes a number of previous 2d-peaks



(a) RT-Intensity plane

(b) RT-m/z plane

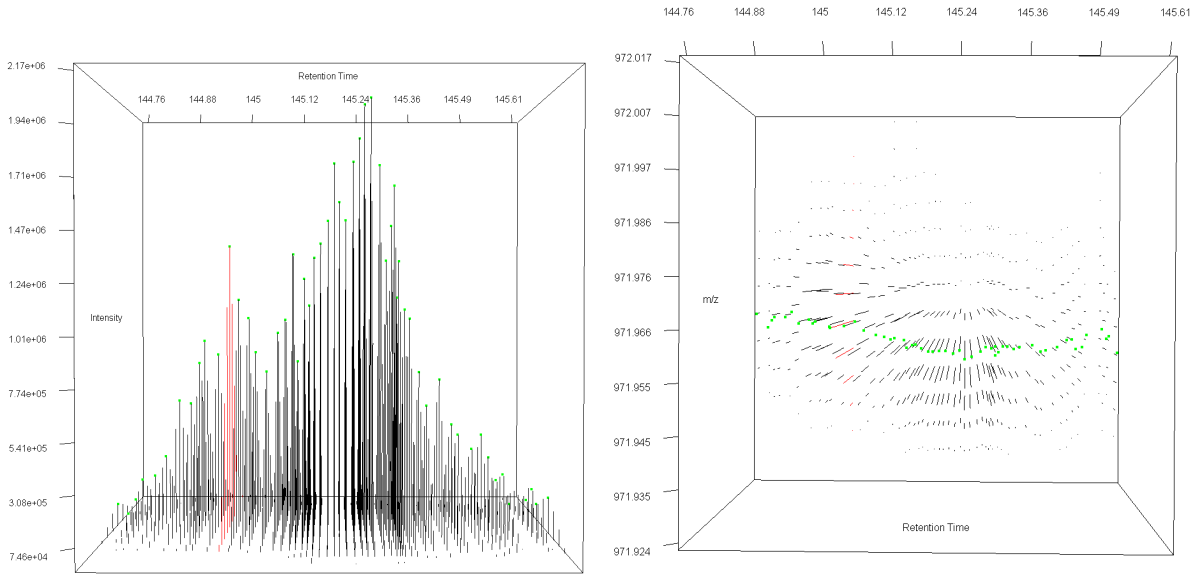
Figure 4.5: Example of 3d-peak, in which a continuous m/z-drift occurs. Note that by inspecting solely the elution profile one can only assume but not with certainty conclude that the peak in fact indicates the presence of two peptides.

(according to the set points to use for the mass drift detection) plus the candidate peak. If in total the m/z-deviation between the consecutive peaks is greater than the tolerance threshold the candidate peak is rejected.

4.3.4 Elution Profile Tracking

There are also situations in which the shape of the elution profile can give a hint where a 3d-peak should end, while the inspection of the RT-m/z plane might not allow this conclusion. Figure 4.6 shows an example of such a peak. Whereas in the elution profile the incision between the two peaks is clearly visible, the m/z-value only deviates slightly between consecutive peaks. The masses of the peptides are too close to each other, s.t. these deviations are well within the expected inaccuracy due to technical noise.

Because of the normal deviations in the elution profile it is not possible to apply a method as



(a) RT-Intensity plane

(b) RT-m/z plane

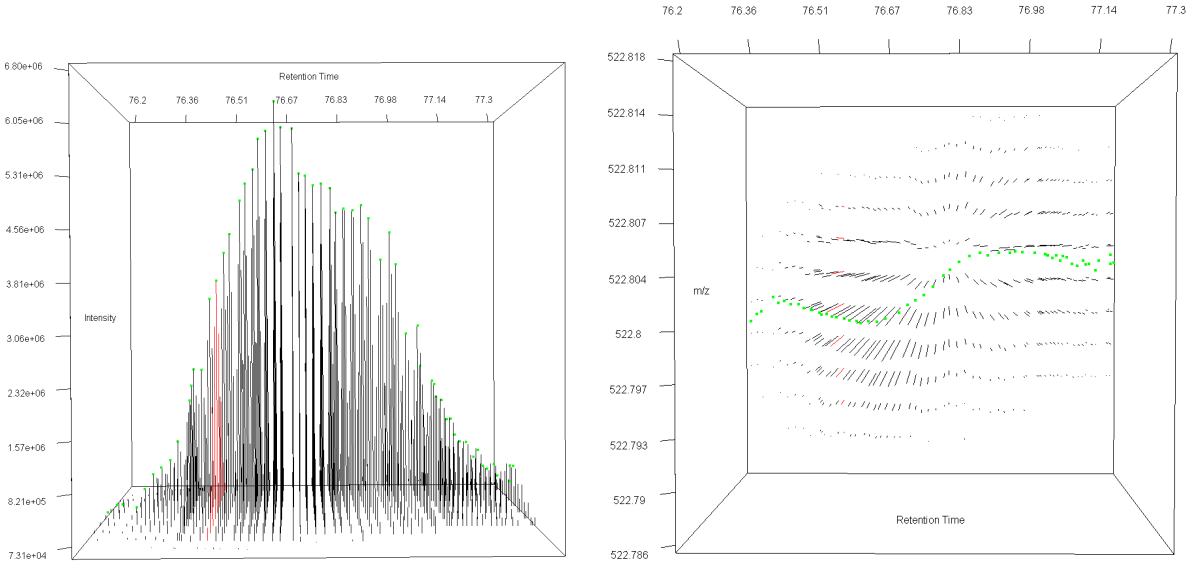
Figure 4.6: Example of 3d-peak split peak. Note that although the incision is obvious in the RT-Intensity plane, the deviations visible in the RT-m/z plane are not conclusive enough to detect it.

simple as in the 2d-peak extraction scenario. The general idea of tracking the slope of the profile is nevertheless a suitable way of detecting the margins of a peak. As an approach capable of compensating these fluctuations we choose to take the intensities of a defined number $n_{reg} - 1$ of previous 2d-peaks, plus the intensity of the candidate peak, into account. Given the retention times RT_i of the 2d-peaks and the and the corresponding intensities I_i, for $1 \leq i \leq n_{reg}$, we perform a linear regression in order to fit a straight line through the data. The slope from the obtained line can then be used as an indicator in the same way as in the 2d-peak setting. I.e. the slope is traced to detect local minima, which most probably indicate the end of the peak.

Finally, to further compensate for possible inaccuracies in the measurements which can lead to unpredictable deviations in the data the algorithm may always skip one spectrum, in which no suitable candidate peak was found or for which the extracted candidate peak did not pass all of the mentioned quality criteria. In this case, the algorithm continues with the next spectrum, as if the skipped one did not exist.
Note that although it is possible to change this setting to allow even more spectra to be skipped, it is not recommended to do so.

4.3.5 Evaluation of the extracted peak

The described quality criteria for the 3d-peak reconstruction are chosen to detect the margins of the peak but not to be too restrictive. However, experiments have shown, that some of the extracted peaks are still obviously incorrect. The algorithm still fails to detect split peaks when the overlap between the two respective peaks is too high. Figure 4.7 shows an example of an obvious split peak which has not been separated by the algorithm. Inspection of the RT-Intensity plane by eye would suggest the conclusion that these are actually two peaks. In the RT-m/z plane the situation is even more obvious. Here the drawback of the decision to design the algorithm in a not too restrictive way becomes apparent.



(a) RT-Intensity plane

(b) RT-m/z plane

Figure 4.7: Example of 3d-peak split peak that could not be detected by the algorithm. Note that although the deviations in both measured dimensions strongly indicate a split peak, the algorithm is too less restrictive to detect it.

Therefore, a final evaluation is done after the peak has been reconstructed. For detecting such split peak the following steps are performed:

- Smooth the n centroid values of the 2d-peaks by applying a 5-point average filter, resulting in $n-4$ smoothed centroids: $c^s_0, \ldots c^s_{n-4}$.
- Excluding c^s_0 and c^s_{n-4}: Determine the range within each 5 consecutive points, i.e. the minimal and maximal m/z value: c^s_{min} and c^s_{max}
- If $c^s_{max} - c^s_{min} \geq 4$ ppm: Reject the 3d-peak altogether.

The idea behind this check is to detect unusual mass shifts in a central area of the peak. In principal, the same concept is followed as for the constant checking of the m/z-deviations during

peak-reconstruction. However, since this check tries to target the central area of the peak, where the m/z-deviations between 2d-peaks should be smallest, it specifically detects the described situation of a split peak occurring right in the middle of the whole reconstructed peak. Moreover, we do not want to target the beginning and the end of the peak, as there the intensities and thus the accuracy of the measured values is lowest and least dependable. The 5-point average filtering, as well as omitting the first and the last of the smoothed m/z-values when performing the check truncate and smooth the data accordingly.

Note further that if this final evaluation of the 3d-peak fails it is rejected completely. In this case, the precursor's original m/z-value which has been provided as input is returned. The same holds for the rare cases where no 2d-peak at all can be found, not even in the original MS1-spectrum. This also simply results in the input value to be returned.

4.4 Further ways of improving the mass precision

Calculating the centroid of a 3d-peak already improves the precision of the precursor-m/z significantly (see section 5.3). But there are of course more possibilities for improvements that can be applied. In this section two further methods, the integration of isotope patterns and linear mass-recalibration, are described.

4.4.1 Usage of Isotope Patterns

As already discussed in section 3.2, we know that each true peak, i.e. each peak stemming from an actual ionized peptide is usually accompanied by several isotope peaks forming an isotope cluster and naturally, such clusters can also be found in MS1-spectra. The information about the other isotopes, besides the original precursor, can and should also be used when determining the precursor m/z-value. Therefore, the 3d-peak reconstruction algorithm is augmented with a per-spectrum isotope detection. I.e. in each spectrum in which a suitable 2d-peak was found and added to the 3d-peak, the algorithm will also search for the isotopes of the 2d-peak found therein. Figure 4.8 illustrates the concept of extracting not one but several peaks from each spectrum.

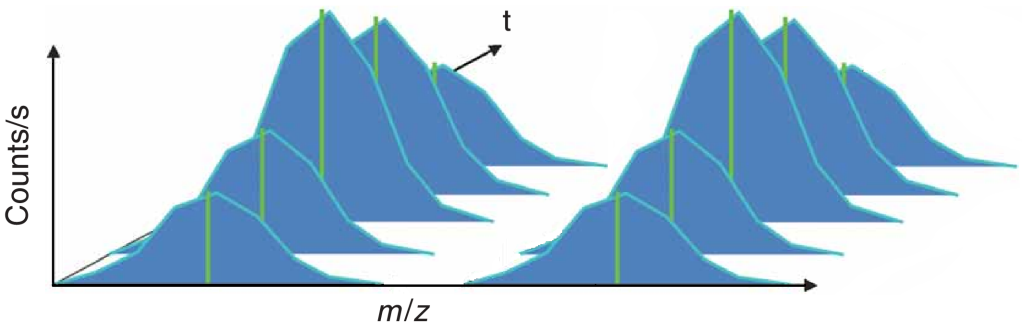


Figure 4.8: Extraction of several peaks per spectrum instead of one. Figure based on the work by Cox et al [7].

Detection of isotopes is done by searching for 2d-peak patterns following the precursor (i.e. the monoisotopic) 2d-peak where the distance between each centroid is ca. $1/z$ and fitting an averagine pattern to evaluate, if the detected pattern is indeed an isotope pattern. The same algorithm that has been developed for the MS2-spectrum manipulation based on the ratios between observed and measured isotope intensities (see section 3.2) can be applied. Note that the charge state of the precursor is known, therefore there is only one possible isotope pattern. The *averagine-test* is therefore rather a check for unperturbed data than if the detected pattern is indeed an isotope pattern. Should the pattern fit well enough, we additionally consider the 1 ^{13}C, 2 ^{13}C and 3 ^{13}C. Further isotope peaks are not considered as they are, if visible at all, too less intense to provide exact uperturbed data. This is due to the usual mass range for peptides in shotgun proteomics experiments, which is between 300 and 3000 Da. See Figure 3.3 in section 3.2, which illustrates these circumstances.

Of course, the obtained centroid values need to be calculated back to the monoisotopic 2d-peak, which is done by simply subtracting a neutron for the 1 ^{13}C isotope, two for the 2 ^{13}C isotope, etc. The centroid values of the additionally acquired 2d-peaks are then also added in the final calculation (eq. 4.19) which computes the centroid by weighted averaging of the 2d-peaks' intensities.
From a statistical point of view the integration of even more 2d-peaks in this equation should further increase the precision of the precursor mass.

4.4.2 Linear Mass Recalibration

The algorithm described so far improves the MS1 precursor precision by extracting and evaluating data from the respective .raw-file and can thus in principle be applied offline and separate from any workflow. The method described in the following needs to be integrated in a workflow. The basic idea to this approach has already been proposed by Cox et al. [7] who perform a recalibration of the precursor masses obtained in the search results. Using high-scoring peptides, an average of their deviations in ppm from the theoretical masses is calculated. Subsequently, this value is subtracted from all the precursor masses.
Cox et al., in fact, also published a much more complex method for recalibration [8]. However, their intention was mainly to replace the lock mass, a *hardware* mechanism used by the instrument to internally correct mass shifts over time. This method basically performs a recalibration of the m/z-values in two dimensions, retention time and m/z.

Since we rely on a working lock mass calibration, which performs the correction of the mass shift over the retention time dimension, we only use one dimension: m/z. Moreover, the manual analysis of several datasets suggested a polynomial of degree 3 as the most suitable fit still avoiding overfitting of the data.
The procedure therefore looks as follows:

1. Perform a search with relaxed search parameters in order to quickly obtain a rough search result.

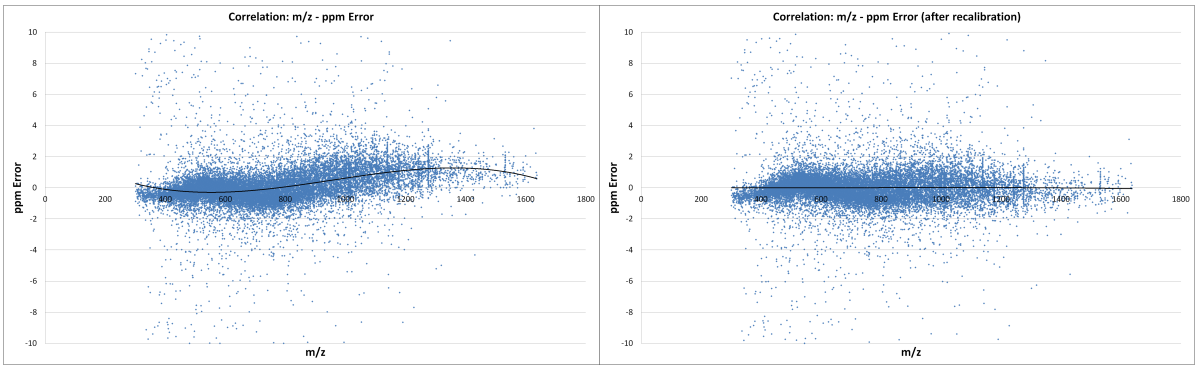
2. Filter the result list for peptides that have a search engine rank of 1, minimum length of 8 and a score that is among the top 5% scores of these peptides.

3. Determine the (m/z, ppm deviation)-pairs for the filtered peptides.

4. Fit a polynomial of degree 3 to this data by applying Linear-Least-Squares minimization.

5. Recalibrate all precursor m/z-values of the search input by subtracting the function value of the polynomial at position m/z from it. This results in a ppm-value distribution around the x-axis, i.e. a normalization of the ppm-errors by setting the average to 0.

Note that the first search has to be done against a normal target database that does not contain any decoy sequences as the recalibration should, of course, be in favor of the target peptides. Otherwise we would introduce a bias towards the decoy peptides.
Figure 4.9 shows the ppm-errors plotted against the m/z-values of high-confident peptides of a sample dataset before and after the recalibration. The benefit of this method is to obtain an average ppm-error of (ideally) 0 in every m/z-range, which in turn also reduces the overall standard deviation. It is therefore, besides the obvious benefit of drawing the average ppm-distribution around 0 and therefore increasing the precision, also necessary to make the results comparable by having an, in a sense, "unbiassed" standard deviation.



(a) ppm-error distribution before recalibration

(b) ppm-error distribution after recalibration

Figure 4.9: Recalibration of precursor m/z-values of a sample dataset to achieve normalization of the ppm-errors around the zero line.

4.5 Resulting Software

The proposed peak-reconstruction algorithm works with very low memory requirements as the MS1-spectra are analyzed one by one. Since the complete algorithm is always run for one specific precursor mass, we can even fairly restrict the range of the MS1-spectra that is loaded into memory to a few Daltons. Furthermore, the applied methods do not require complex calculations, such as the fitting of bivariate Gaussians or similar approaches. We are aware that the runtime could still be improved by an intelligent "alignment" of certain precursors for a simultaneous reconstruction of their elution profiles or even the principal parallelization of the

algorithm. However the implementation of these further features is beyond the scope of this thesis.

The methods described in this section regarding peak-picking were integrated in the *MS2-Spectrum Processor* plugin for the Proteome Discoverer software by Thermo Fisher Scientific. Note, however, that the peak reconstruction can only work, if the full .raw file is provided as an input to the workflow. Without access to the MS1-spectra the method cannot be executed.
Furthermore, the mass recalibration has been incorporated in a second, separate plugin for the Proteome Discoverer, the *Post-Search Mass Recalibrator* node. The recalibration can thus also be applied separately, without precedent peak-reconstruction. Figure 4.10 shows a Proteome Discoverer template, where both nodes are used. First the spectra are selected from the input files, before the peak-reconstruction and/or spectrum cleaning algorithms are applied to them. Afterwards, a search with the Mascot search engine is started. The results are used for the recalibration in the Post Search Mass Recalibrator node, before the spectra with modified precursor masses are submitted to a second search. Afterwards, the results are validated and prepared for visualization.

Since the peak-reconstruction has also been used in other projects, a separate .dll file, *PeakReconstruct.dll*, containing the relevant algorithms has been developed. This library provides functionality for the following additional software products that have been created during the work on this thesis:

1. **Spectrum Analyzer**: A stand-alone tool developed for the in-house analysis of .raw files. The goal of this software is to collect thorough information about the recorded spectra, as well as the important instrument settings. The output provides helpful hints for tuning and optimization of instrument settings for further experiments.
   Moreover, the output is user-definable which helps creating self-tailored analyses of the data to obtain only the relevant information. Here, the library is used to obtain quality criteria of the chromatography, i.e. certain properties of the elution profile of the peptides. These are area, apex retention time, apex intensity and full width at half maximum of the 3d-peaks reconstructed from the provided precursor masses.

2. **PeakAnalyzer**: This program is an internal module of the *SimpatiQCo* software suite developed for the quality control of Thermo mass spectrometers and chromatography columns [36].
   PeakAnalyzer uses the library to obtain the area, apex retention time, apex intensity and full width at half maximum of 3d-peaks reconstructed from the identified peptides that are used for the quality control analysis.

Furthermore, inspired by the work with the .raw files the **SpectrumMerger** has been developed. This software is a stand-alone tool used for consolidating quantification and identification spectra of the same precursor. The idea behind the tool is based on an approach by Köcher et al. who showed that taking two separate spectra for each precursor, one having CID, the other one HCD as activation type, and consecutively merging the reporter ion area of the CID spectrum into the HCD spectrum generates a hybrid spectra combining the benefits of both methods. The generally better quantification achieved by CID is added to the better identification of HCD [22].

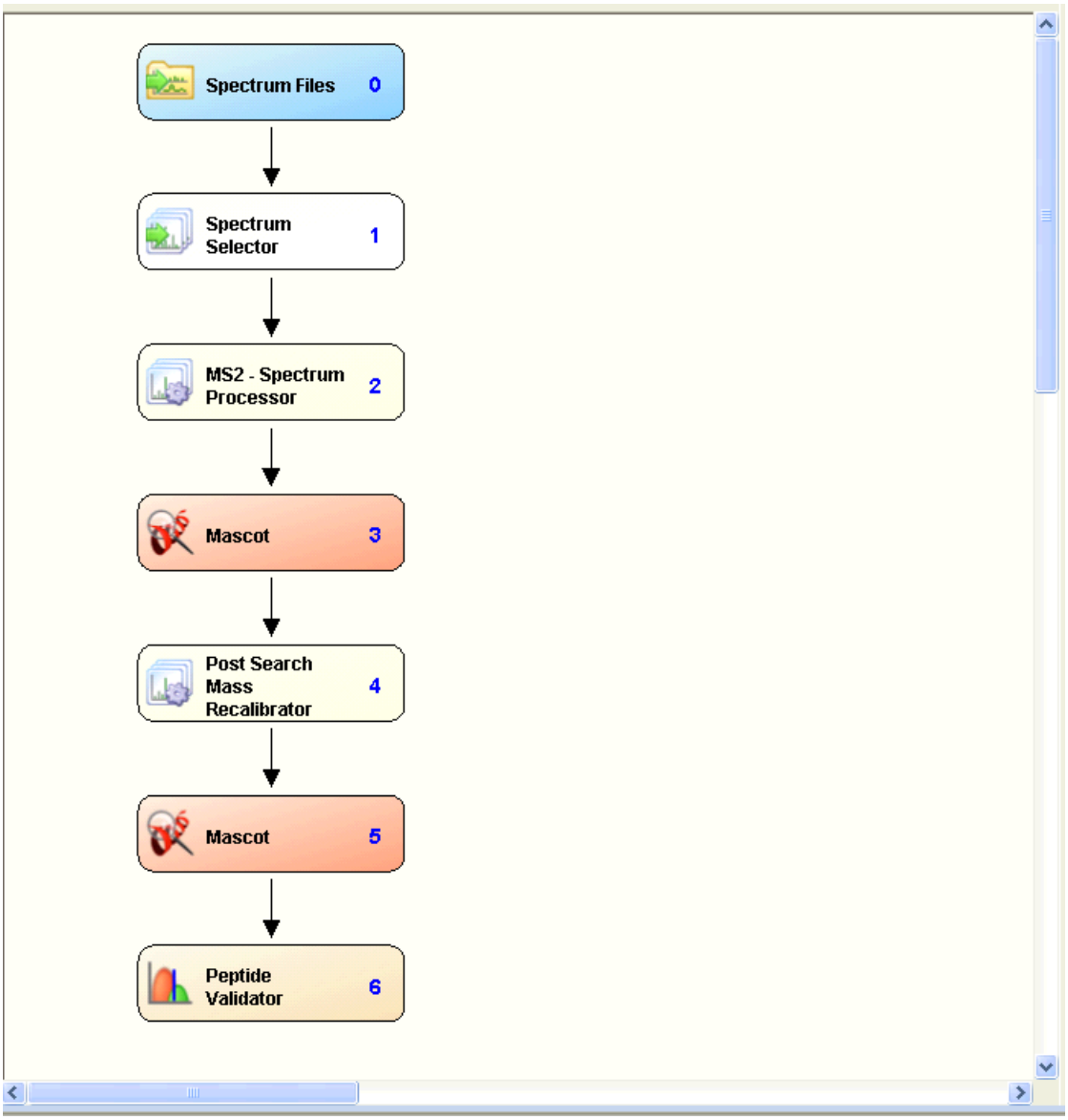


Figure 4.10: A typical Proteome Discoverer (version 1.3.0.339) workflow using both developed nodes: First the input files are defined and the MS2 spectra that should be used from the input are selected. Then, the *MS2 - Spectrum Processor* performs - if desired - deisotoping, charge-deconvolution and precursor mass recalculation. The modified spectra are submitted to a Mascot search whose results are used for recalibration of the spectra by the *Post Search Mass Recalibrator*. The recalibrated spectra are finally submitted to another Mascot search.

This tool was now designed for experimenting with further settings, by allowing automatic merging of two spectra stemming from the same precursor regardless of the specific settings that have been used.

We have showed that the use of not only one but several features from the same and other spectra can drastically improve the precision of precursor masses. Even more precision can be obtained by intelligent recalibration of the precursor masses based on confident identifications of a first search.

Finally, the peak-reconstruction algorithm by itself has proved to be a valuable method for being integrated into instrument-analysis software.

CHAPTER 5

Experimental Results

To test the implemented algorithms we used three different *instances* for the developed algorithms. Since the main focus of this thesis lies on the algorithms developed for improving the precursor m/z-precision, the third instance is a set of three .raw files randomly picked from the 72 files that were used to document the results of the MaxQuant software suite developed by Cox et al [7], on whose idea the algorithm is based.
The instances were chosen to have a certain variety in the samples in order to test the algorithms under different conditions.

Thorough separate analyses of the data were done for both the MS2-Spectrum Manipulation and the Precursor Precision Improvement methods. Finally, an analysis using the complete set of algorithms was conducted.

5.1 Experimental Setup

The following three instances were used for the evaluation of the developed methods:

1. 100ng HeLa, 3h gradient, HCD at NCE 30, isolation width 3 Da, MS1-res. 70000, MS2-res. 17500, recorded by a QExactive Orbitrap instrument.

2. 1μg HeLa, 3h gradient, HCD at NCE 35, isolation width 2 Da, MS1-res. 60000, MS2-res. 7500, recorded by an LTQ Orbitrap Velos instrument.

3. 2μg HeLa, 2h gradient, CID at NCE 35, isolation width 2 Da, MS1-res. 60000, MS2-res. unknown, recorded by an LTQ Orbitrap instrument (3 .raw files).

The samples in all three instances have undergone a tryptic digest before being loaded on the HPLC column. The general setup of search parameters was the following:
All searches were done using the Mascot search engine. It was searched against the *human_swissprot* database, allowing 2 missed cleavage sites. The database was concatenated with a reverse database to allow for a posteriori determination of the FDR.

Unless stated otherwise we used a precursor mass tolerance of 10 ppm and a fragment mass tolerance of 15 mmu (instance 1 and 2) or 0.5 Da (instance 3), respectively.
Possible dynamic modifications were Phospho (ST), Phospho (Y) and Oxidation (M) in the case of instances 1 and 2, and Acetyl (N-term), Oxidation (M), Label:13C(6)15N(2) (K) and Label:13C(6)15N(4) (R) for instance 3. In all cases we added the static modification Carbamidomethyl (C).

All search results were filtered to obtain an FDR of 1% at the PSM level according to the approach described in section 2.2. I.e. the minimum peptide length was set to 8, only results with search engine rank 1 were accepted and the Mascot Ion Score was adapted until the FDR of 1% was reached.

5.2 MS2 - Spectrum Manipulation Results

This section summarizes the observations regarding the MS2 - spectrum manipulation methods that are described in chapter 3 of this thesis. Figures 5.1, 5.2 and 5.3 give a comparison of the PSMs obtained after using the different kinds of deisotoping and deconvolution methods. One can see that the maximum regarding the deisotoping tolerance lies at 15 mmu (for instances 1 and 2, HCD data) and at 80 mmu (for instance 3, CID data). The averagine modeling approach yielded three different results for the three instances: In instance 1 the amount PSMs remained almost the same, in instance 2 we can observe an increase, and for instance 3 the results worsen.

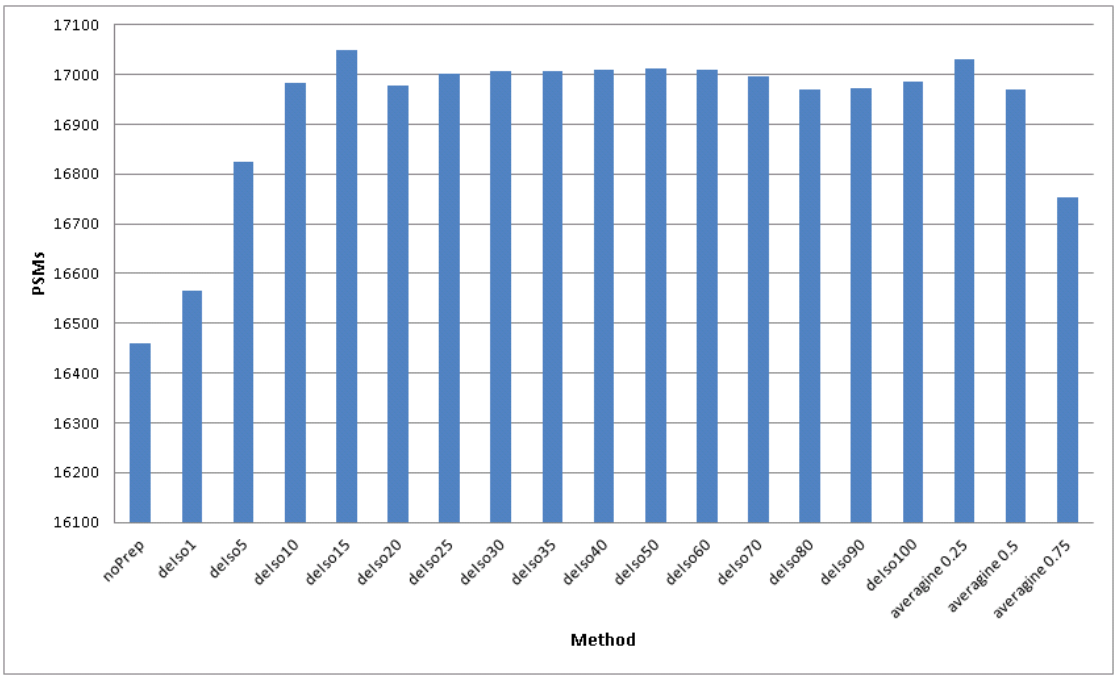
The charge-deconvolution should only be applied if the further processing steps demand it, e.g. if the search engine can only handle singly-charged ions, in this setting its use yields an improvement. If doubly-charged ions are considered as well, the gain by applying charge-deconvolution is almost non-existent.
For the improved charge-deconvolution method we observe that the effect is virtually non-existent. The only difference can be found in instance 1 when comparing both deconvolution approaches after applying deisotoping at 25 mmu tolerance. Only there we see a minimal increase in PSMs.

5.3 Precursor Precision Improvement Results

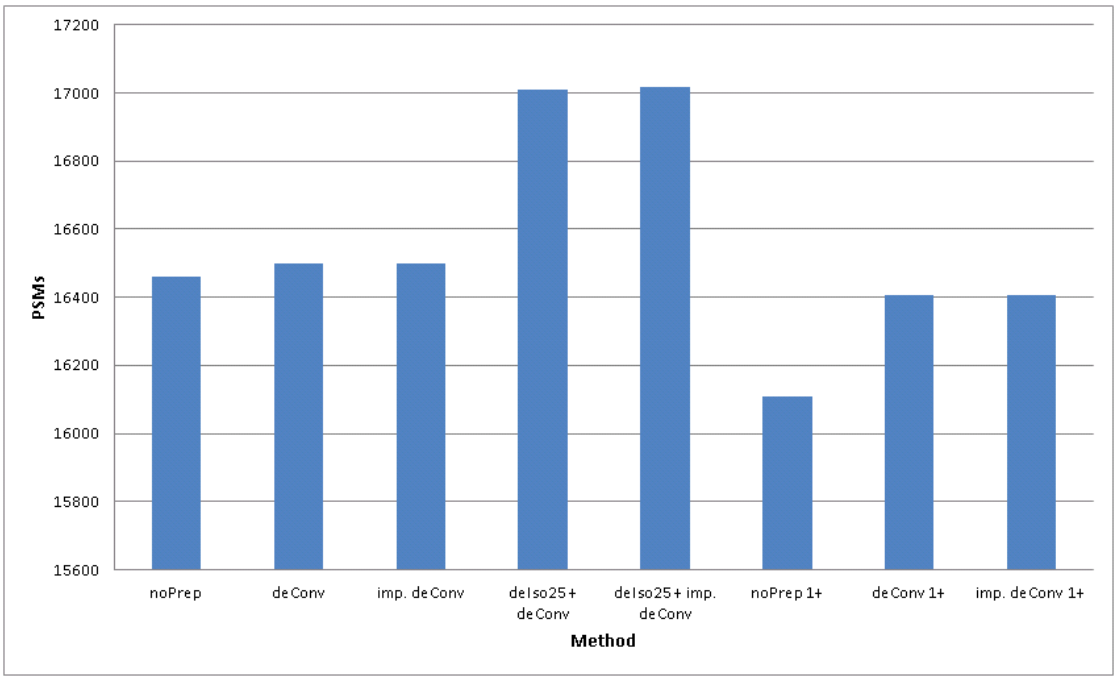
In this section we present the results of the application of the precursor m/z-recalculation algorithm. The first three figures, 5.4, 5.5 and 5.6, show the different ppm-error distributions of the identified peptides for the following three preprocessing methods filtered at 1% FDR and for 10 ppm precursor tolerance:

- without any preprocessing
- with recalculation by peak-reconstruction and using of isotope patterns
- with additionally applying a recalibration of the precursor masses

We observe a significant reduction in the ppm-errors by applying the Precursor Recalculation method, which is expressed by a much narrower and sharper distribution of the ppm-error com-

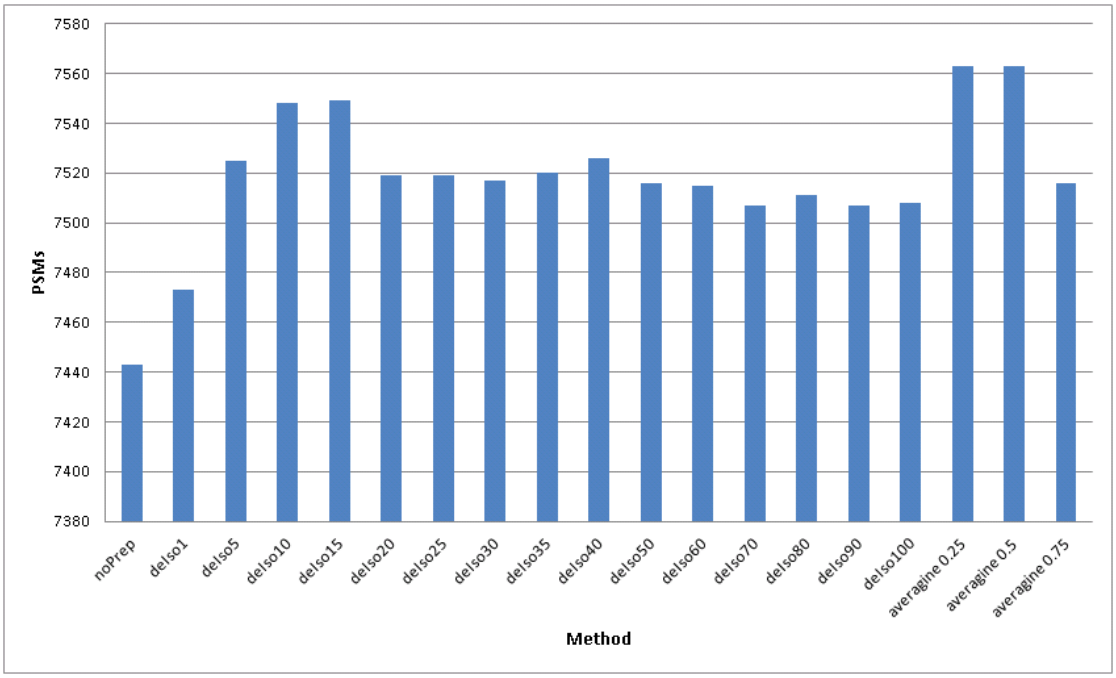


(a) Deisotoping method comparison: Although one can observe a maximum at 15 mmu tolerance used for the deisotoping, using various tolerances up to 100 mmu yields virtually the same reults. Only below 10 mmu does the performance significantly drop. Using the more sophisticated averagine modeling does not yield a significant improvement.

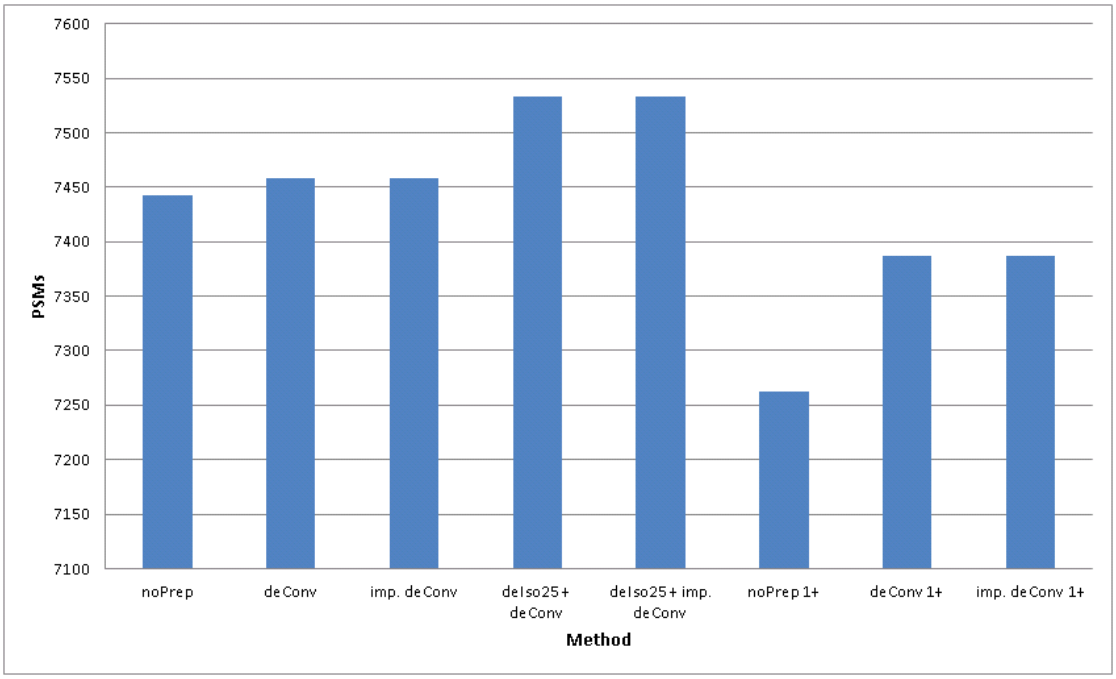


(b) Charge-Deconvolution method comparison: Charge-deconvolution does yield almost no improvement in PSMs. Also when previoiusly applying deisotoping there is no improvement by charge-deconvolution compared to omitting it. The improvement when searching only for singly-charged fragment ions is more significant but in total still worse than including the doubly-charged ions in the search.

Figure 5.1: MS2-spectrum Manipulation method comparison for instance 1.

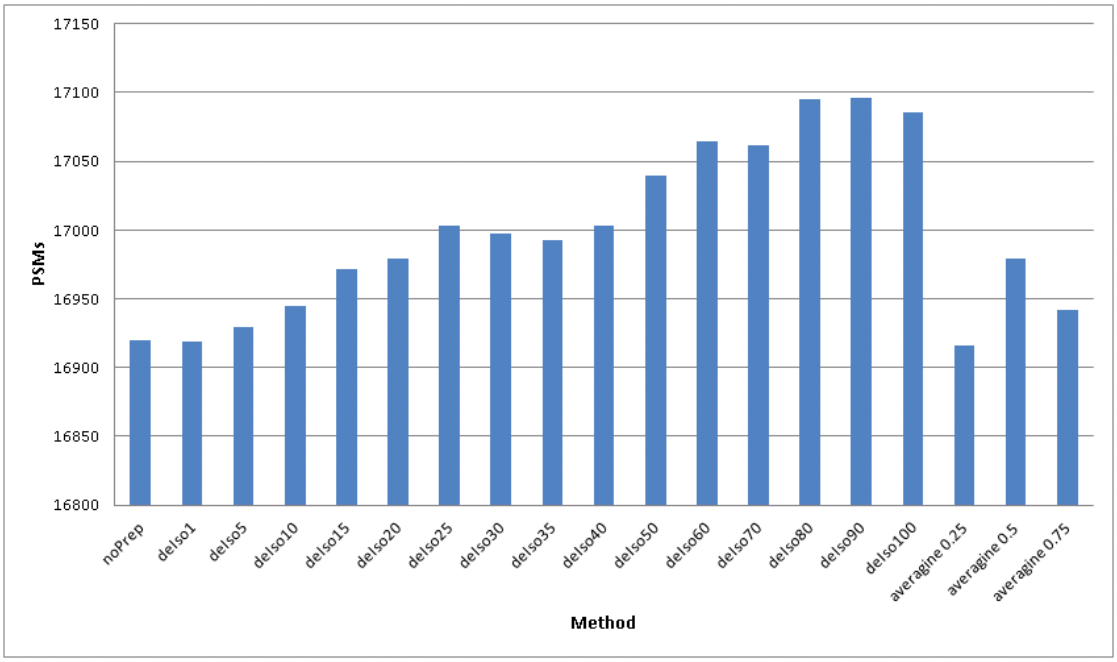


(a) Deisotoping method comparison: Again we see a maximum at 15 mmu deisotoping tolerance. The performance gradually decreases for higher tolerances. For lower tolerances the performance drops much faster. The more sophisticated averagine modeling performed slightly better than the unmodified mehod.

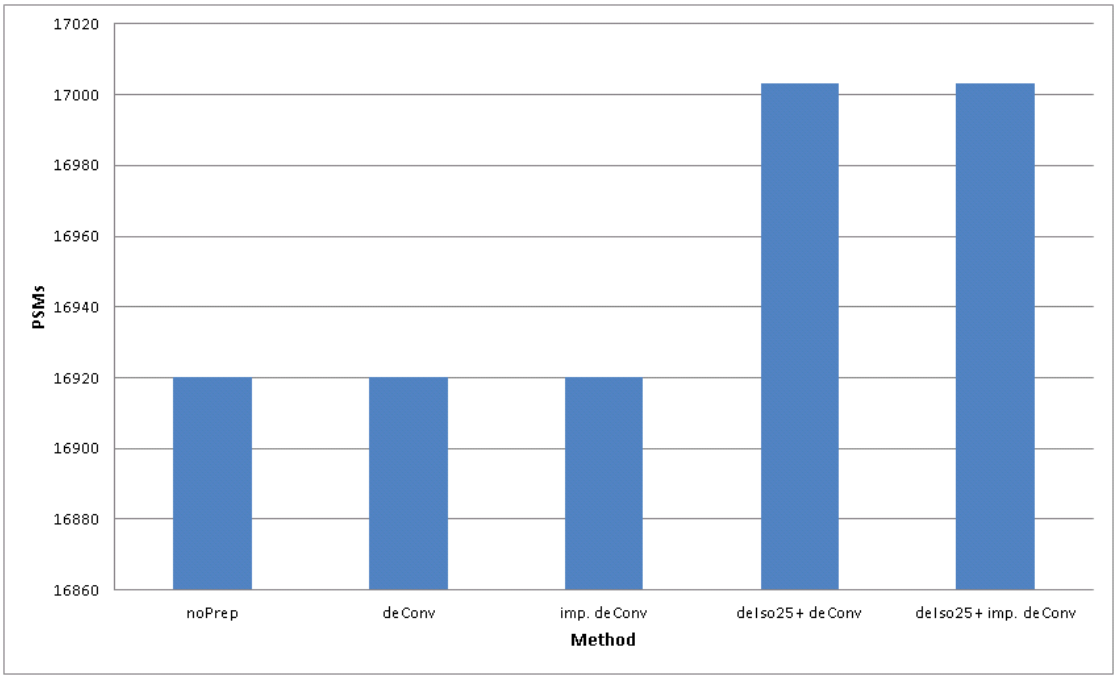


(b) Charge-Deconvolution method comparison: Charge-deconvolution does yield almost no improvement in PSMs. Also when previoiusly applying deisotoping there is no improvement by charge-deconvolution compared to omitting it. The improvement when searching only for singly-charged fragment ions is more significant but in total still worse than including the doubly-charged ions in the search.

 Figure 5.2: MS2-spectrum Manipulation method comparison for instance 2



(a) Deisotoping method comparison: In this case the deisotoping method performed better with increasing tolerance, with a maximum at 90 mmu. In accordance with this trend the more sophisticated averagine modeling performed much worse than the unmodified method.



(b) Charge-Deconvolution method comparison: For this instance charge-deconvolution has no effect regarding PSMs.

Figure 5.3: MS2-spectrum Manipulation method comparison for instance 3

	No Preprocessing	**Precursor Precision Improvement**	**Precursor Precision Improvement & Recalibration**
Average	-0.109	-0.276	-0.002
Median	-0.11	-0.27	-0.02
Std. Dev.	1.633	1.106	1.1
Abs. Average	1.109	0.703	0.63
Abs. Median	0.72	0.46	0.37
Abs. Std. Dev.	1.204	0.897	0.9

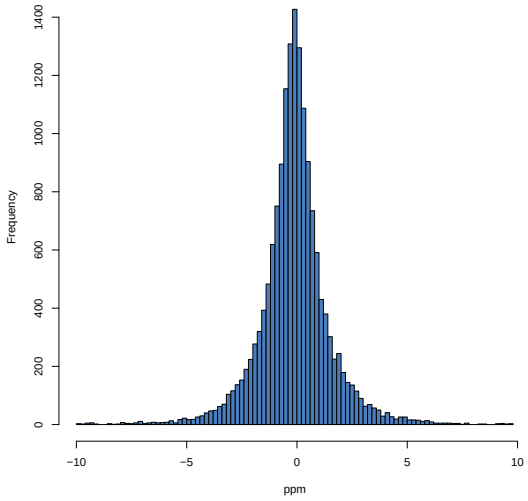
Table 5.1: Comparison of the ppm error-distribution for search instance 1: This table shows the average, median and standard deviation, as well as the absolute average, median and standard deviation for the ppm-error distribution within the identified peptides of instance 1.

	No Preprocessing	**Precursor Precision Improvement**	**Precursor Precision Improvement & Recalibration**
Average	0.042	0.038	-0.01
Median	0.21	0.24	-0.05
Std. Dev.	2.231	1.674	1.448
Abs. Average	1.601	1.059	0.860
Abs. Median	1.12	0.66	0.52
Abs. Std. Dev.	1.554	1.298	1.165

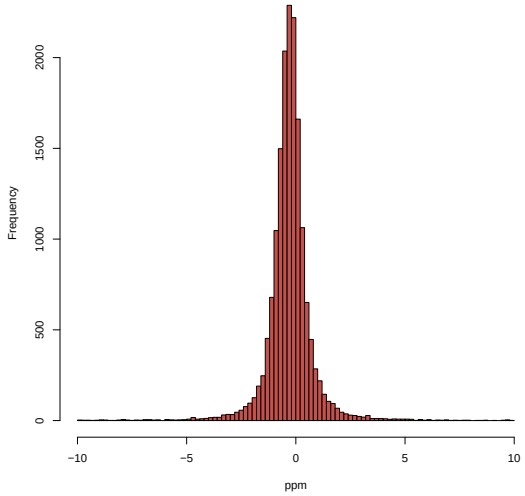
Table 5.2: Comparison of the ppm error-distribution for search instance 1: This table shows the average, median and standard deviation, as well as the absolute average, median and standard deviation for the ppm-error distribution within the identified peptides of instance 2.

pared to the non-preprocessed data. This effect becomes even more pronounced when the precursor m/z-values are additionally recalibrated. Furthermore, Figure 5.5c shows that the bias towards negative ppm values that was present in the data for instance 2 could be removed by the recalibration step.
Another interesting observation is illustrated in Figures 5.4d, 5.5d and 5.6d. Regarding the absolute ppm-errors of the precursor masses after recalculation and recalibration, the figures show that a substantial part of them is below 0.5 ppm. In the case of instance 3, even 84% of the errors were smaller than 0.5 ppm. Tables 5.1, 5.2 and 5.3 list the average, median and standard deviation, as well as the absolute average, absolute median and absolute standard deviation for the three respective instances to clarify the effects the methods have on the ppm-error distribution.

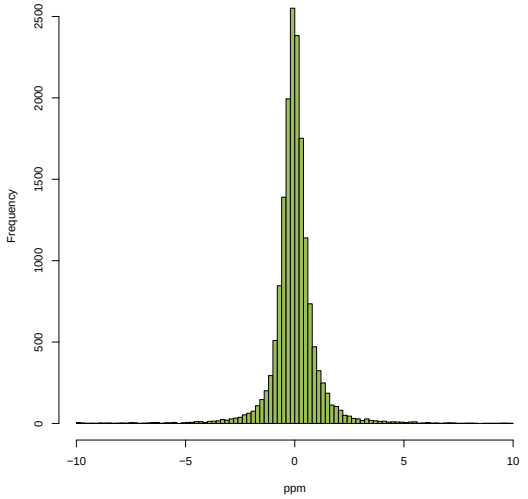
After verifying that there is indeed a benefit in the applied precursor recalculation method, we assessed the search results that can be achieved by using the modified masses. To account for the high precision achieved by this method the search results were additionally filtered for



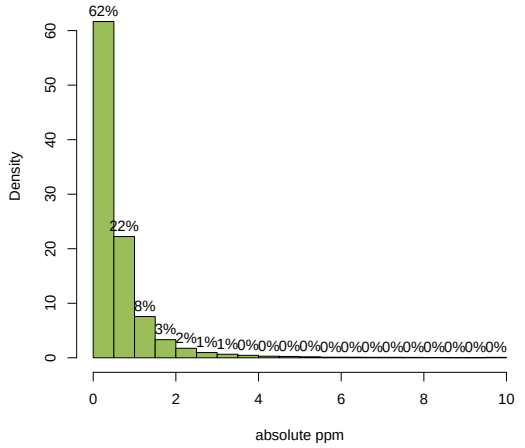
(a) ppm distribution without preprocessing:
average: -0.109, median: -0.11,
standard deviation: 1.633



(b) ppm distribution after precursor recalculation:
average: -0.276, median: -0.27,
standard deviation: 1.106

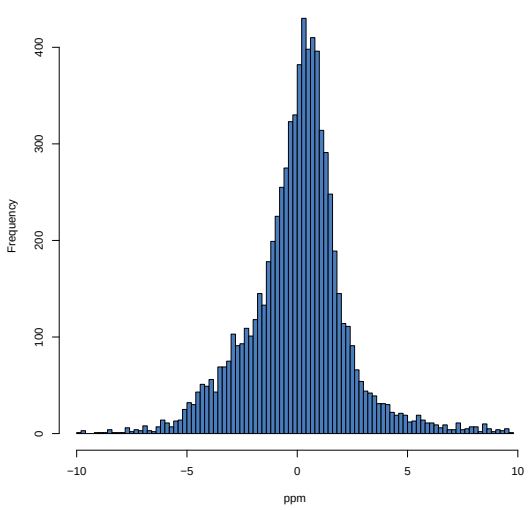


(c) ppm distribution after precursor recalculation
and recalibration:
average: -0.002, median: -0.02,
standard deviation: 1.1

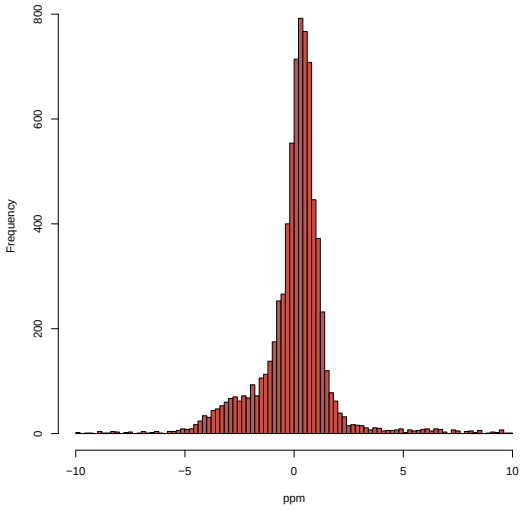


(d) absolute ppm distribution after precursor recalculation
and recalibration:
average: 0.630, median: 0.37,
standard deviation: 0.9

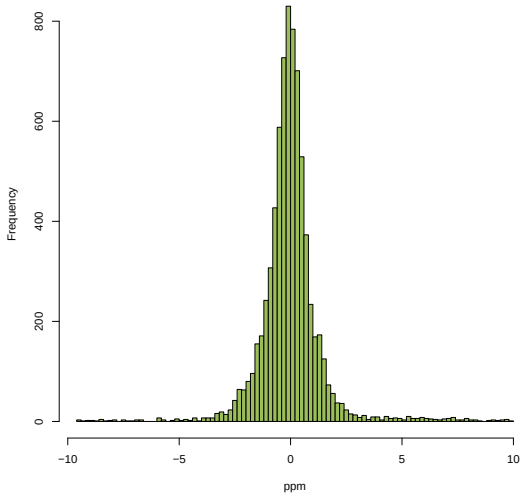
Figure 5.4: ppm-error distribution comparison for search instance 1: The distribution becomes much narrower with a higher peak in the center after recalculation of the precursor masses. This effect is intensified after additional recalibration. After application of both methods 62% of the identified precursors have an absolute mass deviation of less or equal to 0.25 ppm.



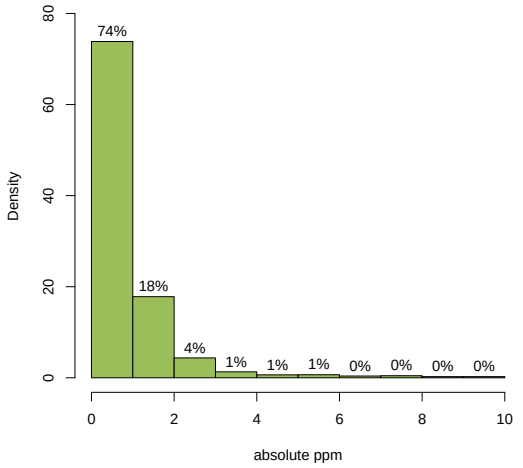
(a) ppm distribution without preprocessing:
average: 0.0418, median: 0.21,
standard deviation: 2.231



(b) ppm distribution after precursor recalculation:
average: 0.038, median: 0.24,
standard deviation: 1.674

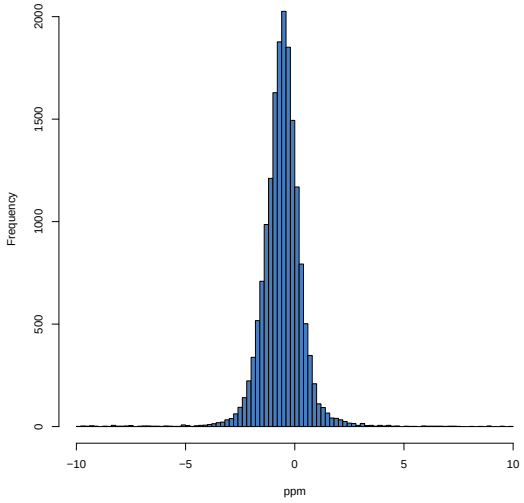


(c) ppm distribution after precursor recalculation
and recalibration:
average: -0.01, median: -0.05,
standard deviation: 1.448

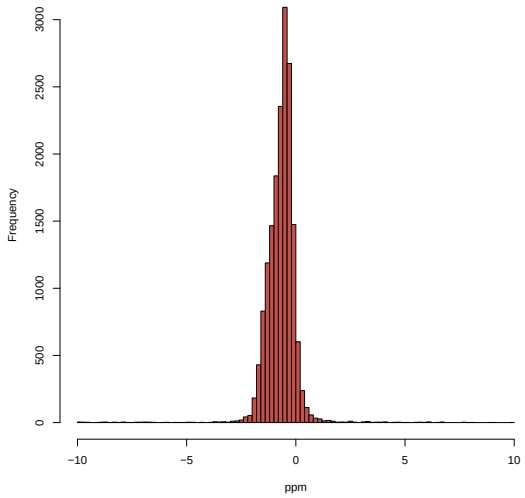


(d) absolute ppm distribution after precursor recalculation
and recalibration:
average: 0.86, median: 0.52,
standard deviation: 1.165

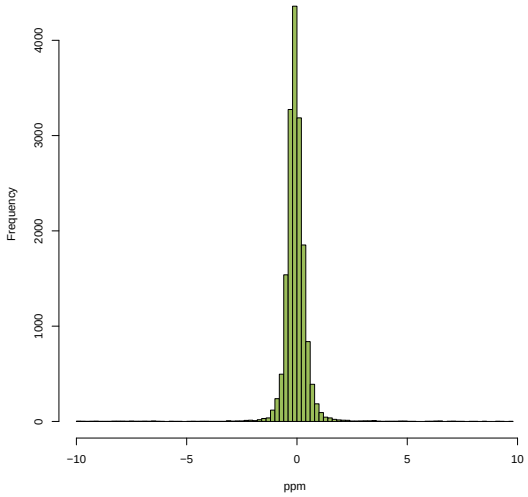
Figure 5.5: ppm-error distribution comparison for search instance 2: The distribution becomes narrower with a higher peak in the center after recalculation of the precursor masses. This effect is intensified after additional recalibration. Moreover, the bias towards negative mass deviations could be corrected by recalibration. After application of both methods 74% of the identified precursors have an absolute mass deviation of less or equal to 0.5 ppm.



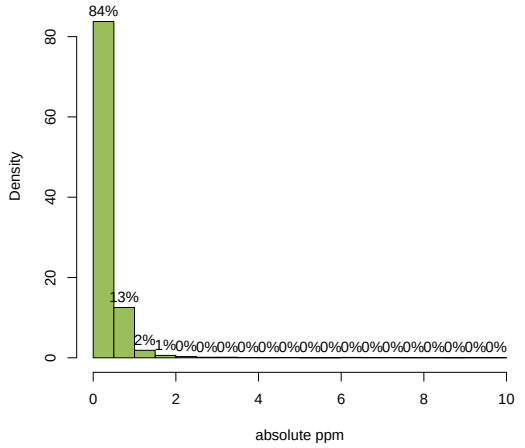
(a) ppm distribution without preprocessing:
average: -0.571, median: -0.55,
standard deviation: 0.996



(b) ppm distribution after precursor recalculation:
average: -0.668, median: -0.6,
standard deviation: 0.7613



(c) ppm distribution after precursor recalculation
and recalibration:
average: -0.059, median: -0.08,
standard deviation: 0.703



(d) absolute ppm distribution after precursor recalculation
and recalibration:
average: 0.346, median: 0.23,
standard deviation: 0.615

Figure 5.6: ppm-error distribution comparison for search instance 3: The distribution becomes much narrower with a higher peak in the center after recalculation of the precursor masses. After additional recalibration the effect increases dramatically. After application of both methods 84% of the identified precursors have an absolute mass deviation of less or equal to 0.25 ppm.

	No Preprocessing	**Precursor Precision Improvement**	**Precursor Precision Improvement & Recalibration**
Average	-0.571	-0.668	-0.059
Median	-0.55	-0.6	-0.08
Std. Dev.	0.996	0.761	0.703
Abs. Average	0.823	0.75	0.346
Abs. Median	0.65	0.61	0.23
Abs. Std. Dev.	0.8	0.681	0.615

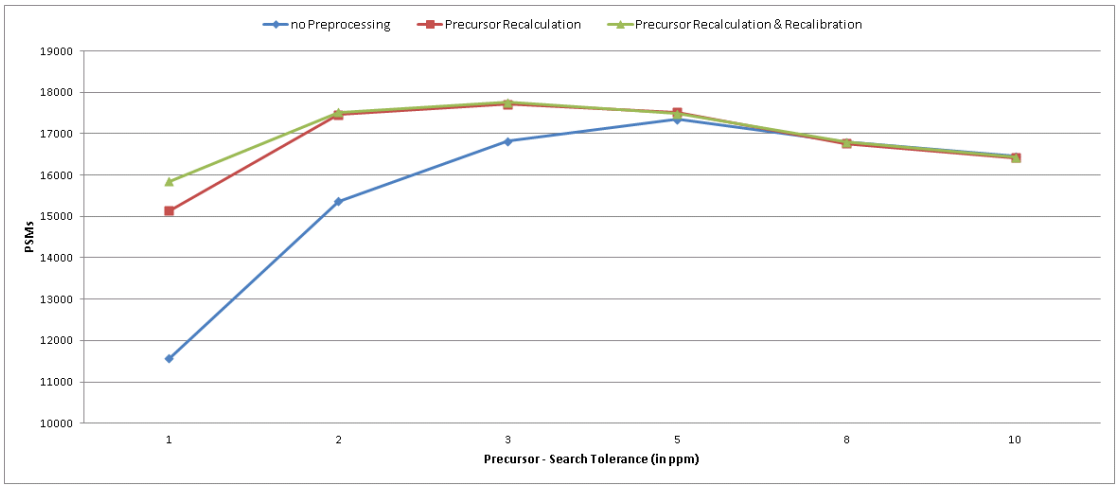
Table 5.3: Comparison of the ppm error-distribution for search instance 1: This table shows the average, median and standard deviation, as well as the absolute average, median and standard deviation for the ppm-error distribution within the identified peptides of instance 3.

different precursor mass deviations, namely 10, 8, 5, 3, 2 and 1 ppm. We compared the resulting PSMs and identified peptides for the different ppm-filter settings, resulting in the curves illustrated in Figures 5.7, 5.8 and 5.9.

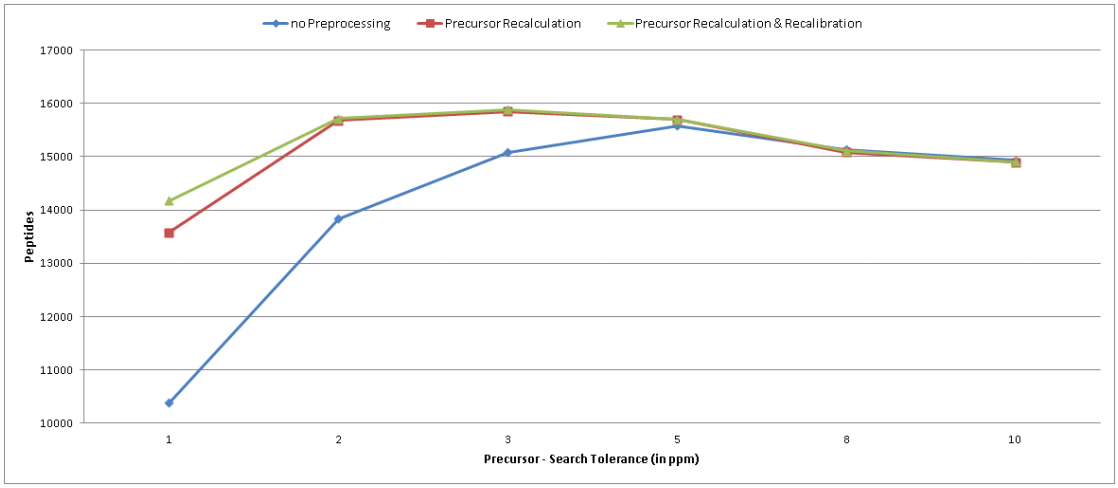
The figures show that the method yields significant improvements for searches with lower ppm tolerances. Moreover, the maxima of both the *Recalculation + Recalibration* approach (green curve), as well as just applying *Recalculation* (red curve) are higher than the maximum of the curve for searches conducted without any preprocessing. I.e. in general one can achieve a moderate increase in identified peptides by applying the method and reducing the precursor tolerance for the search.
However, the increase is not as high as one could have expected by the dramatic increase in precision. This is due to the fact the the Mascot ion score does not punish the distance of precursor masses to the matched peptide. Therefore the benefit of the increased precision only comes into play for small precursor mass tolerances. Tables 5.4, 5.5 and 5.6 show these results in numbers and list for each precursor tolerance the improvement rate the *Recalculation + Recalibration* approach achieved in comparison to the unprocessed spectra.

Finally, we did a comparison between the unprocessed data and the data processed with recalculation & recalibration using the respective ppm tolerance filter settings that yielded the most PSMs. This way, we could compare the best results from both methods in terms of PSMs and identified peptides to assess the actual distance between the most favorable precursor mass filter criteria of both methods. The ppm tolerance filters to select for the three different instances can be read from the mentioned tables, where the best setting in each case is highlighted. Thus, we compared:

- Instance 1: No preprocessing at 5 ppm vs. Recalculation & Recalibration at 3 ppm
- Instance 2: No preprocessing at 8 ppm vs. Recalculation & Recalibration at 3 ppm.
- Instance 3: No preprocessing at 3 ppm vs. Recalculation & Recalibration at 2 ppm.

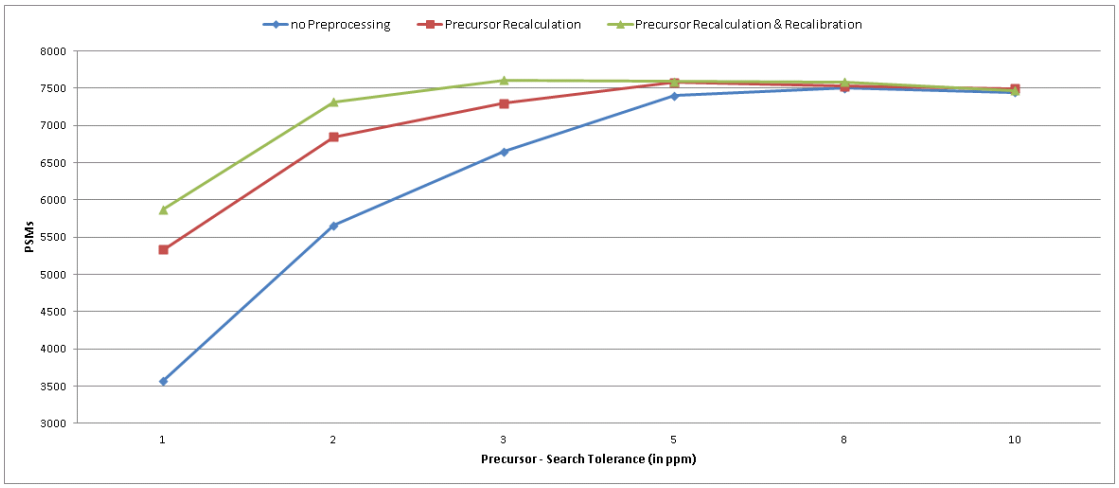


(a) PSMs

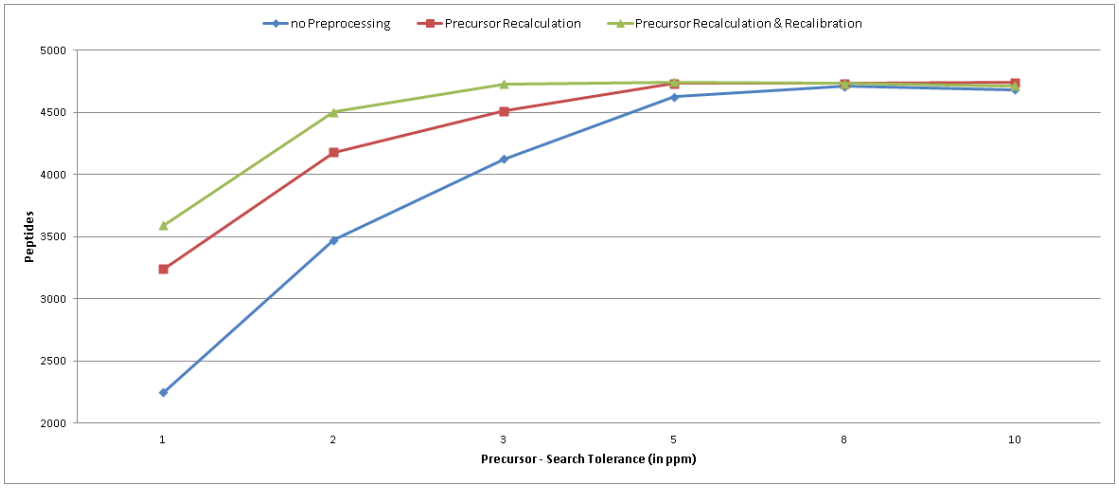


(b) Peptides

Figure 5.7: Search results comparison for search instance 1: For lower precursor-mass deviation filters we observe a dramatic increase in PSMs and indentified peptides when applying the precursor mass recalculation method. Additional recalibration increases this effect. Regarding the maxima of the respectives curves the distance is only moderate.

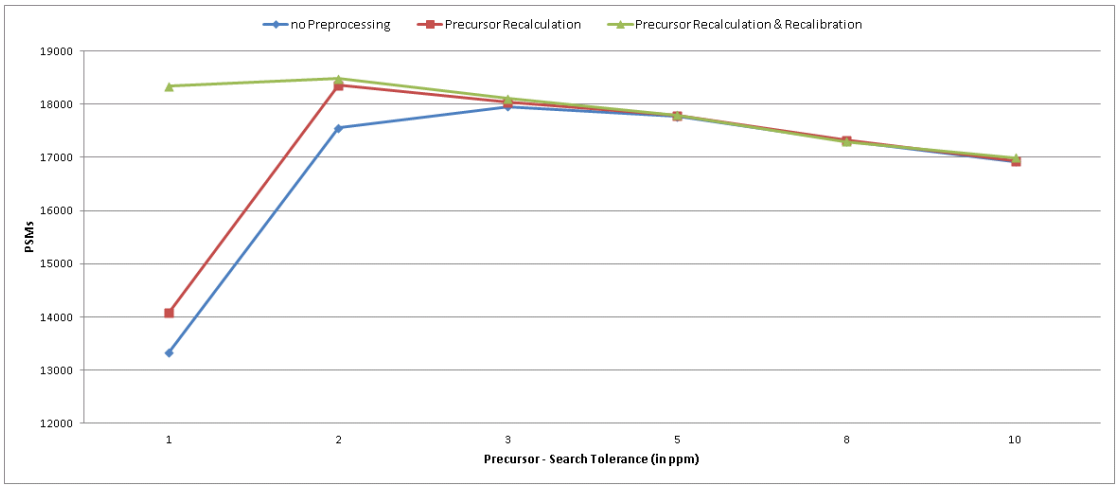


(a) PSMs

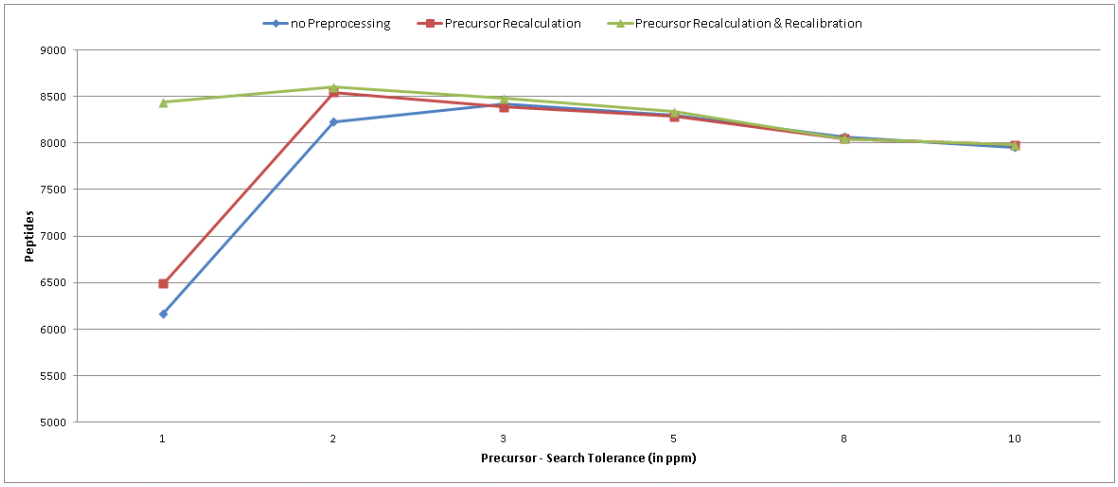


(b) Peptides

Figure 5.8: Search results comparison for search instance 2: For lower precursor-mass deviation filters we observe an increase in PSMs and indentified peptides when applying the precursor mass recalculation method. Additional recalibration further increases this effect significantly. Regarding the maxima of the respectives curves the distance is again only moderate.



(a) PSMs



(b) Peptides

Figure 5.9: Search results comparison for search instance 3: Only for very low precursor-mass deviation filters we observe a significant increase in PSMs and indentified peptides when applying the precursor mass recalculation method. With additional recalibration we observe a dramatic increase for 1 ppm mass deviation. For the other mass deviation settings the performance remains almost the same. Regarding the maxima of the respectives curves the distance is only moderate.

Method	No Prep.	Precursor Prec. Imp.	Precursor Prec. Imp. & Recalibration	Improvement
10 ppm Maximum Precursor Mass Deviation				
PSMs	16458	16425	16436	**-0.13%**
Peptides	14933	14892	14900	**-0.22%**
8 ppm Maximum Precursor Mass Deviation				
PSMs	16794	16758	16792	**-0.01%**
Peptides	15133	15078	15109	**-0.16%**
5 ppm Maximum Precursor Mass Deviation				
PSMs	**17351**	17521	17483	**0.76%**
Peptides	**15578**	15697	15695	**0.75%**
3 ppm Maximum Precursor Mass Deviation				
PSMs	16822	**17717**	**17756**	**5.55%**
Peptides	15083	**15851**	**15882**	**5.30%**
2 ppm Maximum Precursor Mass Deviation				
PSMs	15372	17460	17518	**13.96%**
Peptides	13839	15675	15717	**13.57%**
1 ppm Maximum Precursor Mass Deviation				
PSMs	11567	15142	15845	**36.98%**
Peptides	10383	13579	14178	**36.55%**

Table 5.4: Results comparison for search instance 1: For each method, the highest PSMs- and identified peptides values are written in bold. The last column gives the relative improvement of the precursor precision improvement & recalibration method in comparison to the unprocessed spectra.

The corresponding numbers are listed in table 5.7.

Method	No Prep.	Precursor Prec. Imp.	Precursor Prec. Imp. & Recalibration	Improvement
10 ppm Maximum Precursor Mass Deviation				
PSMs	7443	7497	7467	**0.32%**
Peptides	4682	4738	4711	**0.62%**
8 ppm Maximum Precursor Mass Deviation				
PSMs	**7502**	7528	7581	**1.05%**
Peptides	**4709**	4730	4732	**0.49%**
5 ppm Maximum Precursor Mass Deviation				
PSMs	7402	**7580**	7592	**2.57%**
Peptides	4625	**4733**	**4742**	**2.53%**
3 ppm Maximum Precursor Mass Deviation				
PSMs	6650	7300	**7612**	**14.47%**
Peptides	4124	4510	4726	**14.60%**
2 ppm Maximum Precursor Mass Deviation				
PSMs	5660	6845	7315	**29.24%**
Peptides	3473	4179	4502	**29.63%**
1 ppm Maximum Precursor Mass Deviation				
PSMs	3571	5336	5876	**64.55%**
Peptides	2246	3239	3590	**59.84%**

Table 5.5: Results comparison for search instance 2: For each method, the highest PSMs- and identified peptides values are written in bold. The last column gives the relative improvement of the precursor precision improvement & recalibration method in comparison to the unprocessed spectra.

Method	No Prep.	Precursor Prec. Imp.	Precursor Prec. Imp. & Recalibration	Improvement
10 ppm Maximum Precursor Mass Deviation				
PSMs	16924	16932	16996	**0.43%**
Peptides	7953	7978	7980	**0.34%**
8 ppm Maximum Precursor Mass Deviation				
PSMs	17302	17322	17294	**-0.05%**
Peptides	8069	8045	8047	**-0.27%**
5 ppm Maximum Precursor Mass Deviation				
PSMs	17770	17784	17797	**0.15%**
Peptides	8302	8287	8335	**0.40%**
3 ppm Maximum Precursor Mass Deviation				
PSMs	**17953**	18031	18102	**0.83%**
Peptides	**8424**	8389	8480	**0.66%**
2 ppm Maximum Precursor Mass Deviation				
PSMs	17551	**18354**	**18477**	**5.28%**
Peptides	8228	**8547**	**8603**	**4.56%**
1 ppm Maximum Precursor Mass Deviation				
PSMs	13331	14086	18333	**37.52%**
Peptides	6170	6488	8440	**36.79%**

Table 5.6: Results comparison for search instance 3: For each method, the highest PSMs- and identified peptides values are written in bold. The last column gives the relative improvement of the precursor precision improvement & recalibration method in comparison to the unprocessed spectra.

Instance 1			
	no Preprocessing	**Full method**	**Improvement**
PSMs	17351	18061	**4.09%**
Peptides	15578	16269	**4.44%**
Instance 2			
	no Preprocessing	**Full method**	**Improvement**
PSMs	7502	7676	**2.32%**
Peptides	4709	4764	**1.17%**
Instance 3			
	no Preprocessing	**Full method**	**Improvement**
PSMs	17953	18585	**3.52%**
Peptides	8424	8691	**3.17%**

Table 5.7: Comparison of the results of the full method to the unprocessed spectra for the three instances.

CHAPTER 6

Conclusion and Future Work

In order to cope with the ever-increasing amount of data generated by high-throughput shotgun-proteomics experiments, methods that can order and divide the data have to be developed. One promising approach is the manipulation of recorded MS2-spectra as they contain the relevant data that is sent to the database search engine and are the determining factor of the results of a proteomics experiment. Moreover, the advances in accuracy and resolution of recent instruments allow the use of newer more accurate methods for spectrum manipulation which do apply complex and less accurate techniques that were needed for spectra with lower resolutions.
Therefore, in this thesis preprocessing methods for high resolution, high accuracy proteomics data with the aim of increasing the peptide identification performance of database search engines have been discussed.

We analyzed existing deisotoping and deconvolution approaches for MS2-spectrum manipulation and experimented with possible improvements of the methods. The experiments were conducted using three input instances consisting of a single .raw file and a set of three .raw files, respectively. In summary our experiments showed that only a moderate increase in PSMs can be achieved by these preprocessing steps. Furthermore, we made the interesting observation that neither a refinement nor a degradation of the methods significantly changes this performance increase. This means that there is a balance in the scoring between the removal of as many peaks as possible and the selective removal of a few peaks. The conclusion that can be drawn from this first part of our work is therefore that spectrum cleaning can be useful to gain a few percent more identified peptides. There is, however, not much more benefit to gain when using common search engines such as Mascot, which do not take the mass precision of fragment ions into account, that would justify the implementation of more sophisticated methods than the simple already existing ones.

The second part of the work dealt with the development of an algorithm to improve the precision of the precursor masses associated with the MS2-spectra that are sent to a search engine. We furthermore showed that the recalculation and recalibration methods significantly reduce the average, as well as the variance of ppm-errors in the identified peptides. This highly increases the results for searches with low precursor tolerance. However, the distance between

the best-performing tolerance using these methods and the best-performing tolerance without preprocessing is not as significant. The reason is that common search engines, such as Mascot and SEQUEST, do not consider the distance of precursor ions to the matched peptide when the score for the respective match is calculated.
A new search engine that does take these distances into consideration is currently being developed at the IMP Vienna in cooperation with the University of Applied Sciences Upper Austria and is a starting-point for further experiments and developments of the proposed algorithm. Alternatively, a high-tolerance search followed by a rescoring and reranking of the obtained search results based on the mass tolerance is possible. In any case, it is necessary to account for the precursor precision, since in our opinion it provides additional confidence in the matched peptide, even more, if the mass value is obtained as an average from several spectra following the elution profile of the respective peptide.

Besides the analysis of deisotoping and deconvolution and the development of a reliable method for the increase of precursor mass precision, several software products have resulted from this work. Most importantly, the algorithms were included in a plugin written in C# for the Proteome Discoverer software, version 1.3.0.339, by Thermo Fisher Scientific, that can easily be integrated into existing workflows.
Additionally, the peak-reconstruction algorithm was implemented into two further software products:

- *PeakAnalyzer*: A module for *SimpatiQCo*, a quality control software for LC-MS/MS-systems.
- *Spectrum Analyzer*: A stand-alone tool for the analysis of .raw files.

Of course, in both cases, the MS2-spectrum manipulation, as well as the precursor mass precision improvement, there is still room for further developments. Especially the peak-reconstruction, as well as the recalibration approach still offer room for improvement. E.g. the final evaluation of a reconstructred peak is lacking some kind of repair algorithm, in case a peak is probably not originating from a single peptide. Here statistical approaches, such as the fitting of a bivariate Gaussian could help in determining the part of the peak that should be kept. However, the integration of further enhancements would have gone beyond the scope of this thesis. It is nonetheless our ambition regarding future work to further improve and maintain the developed algorithms.

List of Figures

2.1 Generic structure of an amino acid 6
2.2 Peptide bond formation 7
2.3 N- and C-termine of a peptide 7
2.4 Primary protein structure 8
2.5 Structure levels of a protein 9
2.6 General composition of a mass spectrometer 10
2.7 Scheme of a QExactive Orbitrap instrument 11
2.8 Shotgun-proteomics workflow 13
2.9 Electrospray ionization process 15
2.10 The MS/MS-scan cycle 16
2.11 Possible Cleavage sites in a peptide 18
2.12 Basic workflow of database search engines 20
2.13 Peptide sequence overlap between forward and reverse databases 23
2.14 The target-decoy search workflow 24
2.15 Excerpt of an MS1-spectrum schowing the m/z-values and intensities of some profile points. 27

3.1 Deisotoping of a simple isotope pattern 32
3.2 Overlapping isotope patterns 33
3.3 Probability distribution for the number of ^{13}C-isotopes 36
3.4 ^{12}C isotope ratio comparison 37
3.5 1 ^{13}C isotope ratio comparison 38
3.6 2 ^{13}C isotope ratio comparison 38
3.7 3 ^{13}C isotope ratio comparison 39
3.8 Comparison of Mascot search results, example 1 40
3.9 Comparison of Mascot search results, example 2 40
3.10 Comparison of Mascot search results, example 3 41

4.1 Constructing a 3d-peak from several 2d-peaks 49
4.2 Example of a 2d-split peak 50
4.3 Example of an ideal 3d-peak 54
4.4 Example of a 3d-peak, in which a sudden m/z-shift occurs 56
4.5 Example of 3d-peak, in which a continuous m/z-drift occurs 57

4.6 Example of 3d-peak split peak 58
4.7 Example of 3d-peak split peak that could not be detected 59
4.8 Extraction of several peaks per spectrum instead of one 60
4.9 Recalibration of precursor m/z-values 62
4.10 A typical Proteome Discoverer workflow using both developed nodes 64

5.1 MS2-spectrum Manipulation method comparison for instance 1. 69
5.2 MS2-spectrum Manipulation method comparison for instance 2 70
5.3 MS2-spectrum Manipulation method comparison for instance 3 71
5.4 ppm-error distribution comparison for search instance 1 73
5.5 ppm-error distribution comparison for search instance 2 74
5.6 ppm-error distribution comparison for search instance 3 75
5.7 Search results comparison for search instance 1 77
5.8 Search results comparison for search instance 2 78
5.9 Search results comparison for search instance 3 79

Bibliography

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[2] Robert E. Ardrey. *Liquid Chromatography - Mass Spectrometry: An Introduction*. John Wiley & Sons, Ltd, 2003.

[3] G. Audi and A.H. Wapstra. The 1995 update to the atomic mass evaluation. *Nuclear Physics*, A 595:409–480, 1995.

[4] Sean A Beausoleil, Judit Villén, Scott A Gerber, John Rush, and Steven P Gygi. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology*, 24:1285–1292, 2006.

[5] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. *Biochemistry, Sixth Edition*. W.H. Freeman Company, New York, 2006.

[6] P. De Bievre and P.D.P. Taylor. IUPAC Recommended Isotopic Abundances. *Int. J. Mass Spectrom. Ion Phys.*, 123:149, 1993.

[7] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26:1367–1372, 2008.

[8] Jürgen Cox, Annette Michalski, and Matthias Mann. Software Lock Mass by Two-Dimensional Minimization of Peptide Mass Errors. *J. Am. Soc. Mass Spectrom*, 22:1373–1380, 2011.

[9] Robertson Craig and Ronald C. Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry*, 17:2310–2316, 2003.

[10] Edmond de Hoffmann and Vincent Stroobant. *Mass Spectrometry: Priciples and Applications, Third Edition*. John Wiley & Sons, Ltd, 2007.

[11] Peicheng Du and Ruth Hogue Angeletti. Automatic Deconvolution of Isotope-Resolved Mass Spectra Using Variable Selection and Quantized Peptide Mass Distribution. *Analytical Chemistry*, 78, No. 10:3385–3392, 2006.

[12] Pehr Edman. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chemica Scandinavica*, pages 283–293, 1950.

[13] Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4, 3:207–214, 2007.

[14] Joshua E. Elias and Steven P. Gygi. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. *Methods Mol Biol.*, 604:55–71, 2010.

[15] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates III. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *American Society for Mass Spectrometry*, 5:976–989, 1994.

[16] John B. Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M. Whitehouse. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*, 246:64–71, 1988.

[17] Michael Guilhaus. Principles and Instrumentation in Time-of-flight Mass Spectrometry. *J. Mass Spectrom.*, 30:1519–1532, 1995.

[18] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–925, 2007.

[19] Lukas Käll, John D. Storey, Michael J. MacCoss, and William Stafford Noble. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research*, 7:29–34, 2008.

[20] Michael Karas and Franz Hillenkamp. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. *Analytical Chemistry*, 60 (20):2299–2301, 1988.

[21] Paul J. Kersey, Jorge Duarte, Allyson Williams, Youla Karavidopoulou, Ewan Birney, and Rolf Apweiler. The International Protein Index: An integrated database for proteomics experiments. *Proteomics*, 4:1985–1988, 2004.

[22] Thomas Köcher, Peter Pichler, Michael Schutzbier, Christoph Stingl, Axel Kaul, Nils Teucher, Gerd Hasenfuss, Josef M. Penninger, and Karl Mechtler. High precision quantitative proteomics using iTRAQ on an LTQ Orbitrap: a new mass spectrometric method combining the benefits of all. *Journal of Proteome Research*, 8:4743–4752, 2009.

[23] Matthias Mann, Chin Kai Meng, and John B. Fenn. Interpreting Mass Spectra of Multiply Charged Ions. *Analytical Chemistry*, 61:1702–1708, 1989.

[24] Melissa M. Matzke, Katrina M. Waters, Thomas O. Metz, Jon M. Jacobs, Amy C. Sims, Ralph S. Baric, Joel G. Pounds, and Bobbie-Jo M. Webb-Robertson. Improved quality control processing of peptide-centric LC-MS proteomics data. *Bioinformatics*, 27 no. 20:2866–2872, 2011.

[25] Roger E. Moore, Mary K. Young, and Terry D. Lee. Qscore: An Algorithm for Evaluating SEQUEST Database Search Results. *J. Am. Soc. Mass Spectrom*, 13:378–386, 2002.

[26] Nedim Mujezinovic. *Improved Protein Identification after Fast Elimination of Non-Interpretable Peptide MS/MS Spectra and Noise Reduction*. PhD thesis, Vienna University of Technology, 2007.

[27] Nedim Mujezinovic, Georg Schneider, Michael Wildpaner, Karl Mechtler, and Frank Eisenhaber. Reducing the haystack to find the needle: improved protein identification after fast elimination of non-interpretable peptide MS/MS spectra and noise reduction. *BMC Genomics*, I I, Supplement 1, 2010.

[28] Sven Nahnsen, Andreas Bertsch, Jörg Rahnenführer, Alfred Nordheim, and Oliver Kohlbacher. Probabilistic Consensus Scoring Improves Tandem Mass Spectrometry Peptide Identification. *Journal of Proteome Research*, 10:3332–3343, 2011.

[29] National High Magnetic Field Laboratory. http://www.magnet.fsu.edu/. Accessed: 2012-31-08.

[30] National Human Genome Research Institute. http://www.genome.gov. Accessed: 2012-06-20.

[31] Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4, 9:709–712, 2007.

[32] Jesper V. Olsen, Shao-En Ong, and Matthias Mann. Trypsin cleaves exclusively C-terminal to Arginine and Lysine Residues. *Molecular & Cellular Proteomics*, 3.6:608–614, 2004.

[33] Akhilesh Pandey and Matthias Mann. Proteomics to study genes and genomes. *Nature*, 405:837–846, 2000.

[34] DJC Pappin, D Rahman, HF Hansen, M Bartlet-Jones, W Jeffery, and AJ Bleasby. Chemistry, mass spectrometry and peptide-mass databases: Evolution of methods for the rapid identification and mapping of cellular proteins. *Mass Spectrometry in the Biological Sciences*, pages 135–150, 1996.

[35] David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.

[36] Peter Pichler, Michael Mazanek, Frederico Dusberger, Lisa Weilnböck, Christian Huber, Christoph Stingl, Theo Luider, Werner L Straube, Thomas Köcher, and Karl Mechtler. SIMPATIQCO (SIMPle AuTomatIc Quality COntrol): A Server-Based Software Suite which Facilitates Monitoring the Time Course of LC-MS Performance Metrics on Orbitrap Instruments. accepted: 17-09-2012.

[37] P. Roepstorff and J. Fohlman. Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomedical Mass Spectrometry*, 11, 11:601, 1984.

[38] Mikhail M Savitski, Toby Mathieson, Isabelle Becher, and Marcus Bantscheff. H-Score, a Mass Accuracy Driven Rescoring Approach for Improved Peptide Identification in Modification Rich Samples. *Journal of Proteome Research*, 9 (11):5511–5516, 2010.

[39] Michael W. Senko, Steven C. Beu, and Fred W. McLafferty. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *American Society for Mass Spectrometry*, 6:229–233, 1995.

[40] Youting Sun, Jianqiu Zhang, Ulisses Braga-Neto, and Edward R Dougherty. BPDA - A Bayesian peptide detection algorithm for mass spectrometry. *BMC Bioinformatics*, 11:490, 2010.

[41] Danielle L. Swaney, Graeme C. McAlister, Matthew Wirtala, Jae C. Schwartz, John E.P. Syka, and Joshua J. Coon. A supplemental activation method for high efficiency electron transfer dissociation of doubly protonated peptide precursors. *Analytical Chemistry*, 79:477–485, 2007.

[42] Selene K. Swanson and Michael P. Washburn. The continuing evolution of Shotgun Proteomics. *Drug Discovery Today: Targets*, 10:719–725, 2005.

[43] John E. P. Syka, Joshua J. Coon, Melanie J. Schroeder, Jeffrey Shabanowitz, and Donald F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.*, 101, 26:9528–9533, 2004.

[44] Thomas Taus. phosphoRS: A novel probability-based algorithm for sensitive protein phosphorylation-site localization from high-throughput LC-MS/MS data, Vienna University of Technology, 2010.

[45] The Wikimedia Foundation. http://en.wikipedia.org. Accessed: 2012-31-08.

[46] Harry Towbin, Theophil Staehelin, and Julian Gordon. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc Natl Acad Sci USA*, 76, No.9:4350–4354, 1979.

[47] Michael P. Washburn, Dirk Wolters, and John R. Yates III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19:242–247, 2001.

[48] Valerie C. Wasinger, Stuart J. Cordwell, Anne Cerpa-Poljak, Jun X. Yan, Andrew A. Gooley, Marc R. Wilkins, Mark W. Duncan, Ray Harris, Keith L. Williams, and Dr. Ian Humphery-Smith. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *ELECTROPHORESIS*, 16:1090–1094, 1995.

[49] Cathy H. Wu, Lai-Su L. Yeh, Hongzhan Huang, Leslie Arminski, Jorge Castro-Alvear, Yongxing Chen, Zhangzhi Hu, Panagiotis Kourtesis, Robert S. Ledley, Baris E. Suzek, C.R. Vinayaka, Jian Zhang, and Winona C. Barker. The Protein Information Resource. *Nucleic Acids Research*, 31, 1:345–347, 2003.

[50] Vicki H. Wysocki, George Tsaprailis, Lori L. Smith, and Linda A. Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.*, 35:1399–1406, 2000.

[51] Ying Zhang, Zhihui Wen, Michael P. Washburn, and Laurence Florens. Improving Proteomics Mass Accuracy by Dynamic Offline Lock Mass. *Analytical Chemistry*, 83 (24):9344–9351, 2011.

[52] Roman A. Zubarev, Neil L. Kelleher, and Fred W. McLafferty. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *J. Am. Chem. Soc.*, 120:3265–3266, 1998.