TECHNISCHE
UNIVERSITÄT
WIEN

VIENNA
UNIVERSITY OF
TECHNOLOGY

D I S S E R T A T I O N

Improved Protein Identification after Fast Elimination of Non-Interpretable Peptide MS/MS Spectra and Noise Reduction

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Univ.-Prof. Dipl.-Ing. Dr. Günther Raidl
Institut für Computergraphik und Algorithmen E186
Technische Universität Wien

und wesentlicher Mitbetreuung von

Dr.rer.nat. Dr.habil. Frank Eisenhaber
Forschungsinstitut für Molekulare Pathologie Wien

eingereicht an der Technischen Universität Wien
Fakultät für Informatik

von

Nedim Mujezinovic
Matrikelnummer 9726485
Eckertgasse 16/16, 1100 Wien

Wien, am

Nedim Mujezinovic

Zusammenfassung

Tandem-Massenspektrometrie (MS/MS) ist die Standardmethode für die Proteinidentifikation in biologischen Präparaten. In Proteomics-Studien behindert aber die große Zahl der zu bearbeitenden MS/MS-Spektren und deren Kontaminierung mit Hintergrund-Peaks die schnelle und zuverlässige computergestützte Interpretation. Typischerweise tragen weniger als 1% der Spektren pro Präparat und nur etwa 10% der Peaks pro Spektrum zum Endresultat bei. Die Hintergrund-Peaks in den Spektren stammen nicht nur von den Isotopenvarianten und mehrfach geladenen Replikaten der Peptid-Fragmentationsprodukte, sondern auch von unbekannten Fragmentationswegen, präparatspezifischen oder systematischen chemischen Kontaminationen oder vom Rauschen der empfindlichen elektronischen Nachweissysteme. Neben der dramatischen Verlängerung der Rechenzeit der Interpretationssoftware kann der Hintergrund auch zur falschen Proteinidentifikation führen, insbesondere bei *de novo*- Sequenzierungsalgorithmen.

In dieser Arbeit wurden unter anderem zwei schnelle Verfahren entwickelt, die den "Heuhaufen" der MS/MS-Daten wesentlich reduzieren: (1) Sequenzleiterregeln sortieren Spektren aus, von denen sich keine Peptidsequenzen ableiten lassen. (2) Techniken auf Basis Modifizierter Fourier-Transformation löschen einen Teil des Hintergrunds in den verbleibenden Spektren. Im Durchschnitt müssen nur ca. 35% der ursprünglichen MS/MS-Spektren, die wiederum um ca. ein Viertel in ihrer Größe reduziert wurden, an die Interpretationssoftware übergeben werden. Dies wird faktisch ohne Verlust an Information und mit einer erhöhten Sequenzabdeckung erreicht, obwohl die benötigte Rechenzeit um etwa zwei Drittel reduziert wurde. Der Algorithmus wurde in Form der Anwendung *MS Cleaner* implementiert.

Abstract

Tandem mass spectrometry (MS/MS) has become a standard method for protein identification in biological samples, but in large-scale proteomics studies, the huge number and the noise contamination of MS/MS spectra obstruct swift and reliable computer-aided interpretation. Typically, less than 1% of the spectra per sample and about 10% of the peaks per spectrum contribute to the final result. The background peaks in the spectra result not only from isotope variants and multiply charged replicates of the peptide fragmentation products but also from unknown fragmentation pathways, sample-specific or systematic chemical contaminations or from noise generated by the electronic detection system. Besides dramatically prolonged computation time, the noise can lead to incorrect protein identification, especially in the case of *de novo* sequencing algorithms.

Two fast screens can essentially reduce the haystack of MS/MS data: (1) Sequence ladder rules remove spectra non-interpretable in peptide sequences. (2) Modified Fourier-transform-based criteria clear background in the remaining data. On average, only a rest of 35% of the MS/MS spectra (each reduced in size by about one quarter) have to be handed over to the interpretation software with proportional decrease of computer resource consumption, essentially without loss of information and a trend to improved sequence coverage.

In this work, an algorithm for detection and transformation of multiply charged peaks into singly charged monoisotopic peaks, removal of heavy isotope replicates and random noise is described. The approach is based on numerical spectral analysis and signal detection methods. The algorithm has been implemented in a stand-alone computer program called *MS Cleaner*.

Acknowledgments

First of all I thank Dr. Frank Eisenhaber of the Bioinformatics Group at the IMP for his consistent support and advice throughout the whole project. Also I would like to thank Prof. Günther Raidl of the University of Technology Vienna for his encouragement and especially for helpful advice regarding algorithms. I am grateful to Karl Mechtler of the Protein Chemistry Facility of the IMP for providing me with the opportunity to work on this stimulating project.

My thanks also go to my IMP colleagues Dr. Maria Novatchkova, Dr. James Hutchins, Georg Schneider, Georg Kraml, Michael Wildpaner, Christoph Stingl and Ines Steinmacher for many helpful contributions during this work. I am also grateful to Thomas Burkard, Alex Schleiffer, Birgit Eisenhaber, Georg Neuberger, Werner Kubina, Tian Sun, Christian Brandstätter and members of the Protein Chemistry Facility for a productive working atmosphere.

I would like to thank Tarik Mehmedovic, Abdulkadir Hasanagic, Sead Grebovic and Edin Ibrisimovic for being such good friends, and especially Mehmedalija Mutapcic, Hazim Cebic and Damir Ibrisimovic for their deep, honest and loyal friendship.

Finally, I would like to express my enormous gratitude to my father Muhidin, my mother Suada and my sister Vildana. Their constant moral support, help, encouragement and love have been an inspiration throughout my life. I hope they are aware of how grateful I am to them.

List of Abbreviations

AA	Amino Acids
ADH	Yeast Alcoholdehydrogenase
API	Atmospheric Pressure Ionization
BSA	Bovine Serum Albumine
CE	Capillary Electrophoresis
CID	Collision Induced Dissociation
DC	Direct Current (voltage)
DLL	Dynamically Linked Library
GC	Gas Chromatography
HPLC	High Pressure Liquid Chromatography
IMP	Research Institute of Molecular Pathology Vienna
LC	Liquid Chromatography
MALDI	Matrix Assisted Laser Desorption Ionization
MS	Mass Spectrometry
PVM	Parallel Virtual Machine
QTOF	Quadrupole Time Of Flight
RF	Radio Frequency
TOF	Time of Flight
TRF	Human Transferrin

Contents

Chapter 1

Introduction

Developments in modern mass spectrometry (MS) made possible the large-scale analysis of cellular proteomes [1, 2, 3]. Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) is the standard technique used for analysis of complex protein mixtures [4, 5]. Since modern mass spectrometers can generate large data sets with high throughput, computational analysis of thousands of spectra has become the major bottleneck. The accuracy of the computer-generated interpretations (the identity of the proteins and their post-translational modifications) as well as the time and the storage requirements for their computation are highly dependent on the quality of MS/MS spectra. As a measurement for the quality of spectra, the existence of peaks that support the fragmentation model of real processes in mass spectrometer, as well as a desirably small number of non-interpretable peaks are the main criteria.

In many cases, but not always, b- and y-ions and their derivatives resulting from cleavage at peptide bonds are the most dominant signals in MS/MS spectra of peptides after their fragmentation by low-energy collision-induced dissociation (CID) [5, 6, 7, 8, 9, 10, 11, 12]. However, MS/MS spectra typi-

cally contain many more peaks than can be expected from this fragmentation scheme. Typically, less than 1% of the spectra per sample and about 10% of the peaks per spectrum contribute to the final result if the noise does not even prevent protein identification. Some of the peaks are repeated shifted signals due to the natural isotope distribution [13]. The heavy isotope variants and the monoisotopic peak form isotope peak clusters that can be detected with high-resolution instruments. Electrospray ionization (ESI) allows measuring the masses of large molecules by producing multiply charged ions, thereby decreasing the mass-over-charge ratio into detectable ranges [14, 15, 16, 17, 18]. If a fragment ion comprises several functional groups capable of acting as a charge carrier, the same isotope peak cluster can be repeated with a different charge state at different mass-over-charge values in the spectrum. Other signals originate from unknown fragmentation pathways, sample-specific or systematic chemical contaminations and random noise produced by the electronic detection system.

It is hardly possible to derive any benefit from the above mentioned additional background peaks that can compose the majority of the spectrum as long as the theoretical understanding of the mechanism of their genesis is scarce. The presence of these peaks does not only complicate computer-based spectrum interpretation by increasing the computation time. More critically, false interpretation of high-intensity signals as potential b- or y-related ions can lead, in some cases, to incorrect sequence interpretations of proteins or false identification of their post-translational modifications. Particularly, the de novo sequencing approach [19, 20, 21, 22, 23, 24, 25] is affected by this problem, where each peak is part of a sequence puzzle to be solved and, therefore, has initially to be considered as a potential b- or y-ion. In the case of algorithms based on protein sequence database searches

[26, 27, 28, 29, 30, 31, 32], the danger of misinterpretation is not so dramatic, especially for protein targets without post-translational modifications, since the space of naturally occurring protein sequences is much smaller than the set of sequences that can be theoretically generated. Usually, few dominating peaks originating from the fragmentation along peptide bond are sufficient to unambiguously determine the register of a peptide fragment within the original protein sequence. But when the nature of possible post-translational modifications is *a priori* unknown (and, therefore, the mass changes to be anticipated vary widely) or when the database contains many proteins with similar peptides, the background can lead database search methods down a wrong path and result in incorrect protein identification.

In this work, I propose solutions for these questions and emphasize the benefits of pre-processing and cleaning of MS/MS spectra. For this purpose, new algorithms and methods were developed for deisotoping, deconvolution (recognition of multiply charged peak clusters), random noise removal and detection of non-interpretable spectra. These deisotoping and deconvolution algorithms are capable of finding singly and multiply charged isotope cluster even if MS/MS spectra do not show clear isotope distribution. For this purpose, MS/MS spectra have been investigated both in the mass-to-charge coordinate and in the Fourier-transformed frequency domain. The deisotoping procedure was performed by applying signal processing filters in the frequency domain. Correlation analysis of experimental MS/MS signals with theoretical calculated isotope patterns has been shown as a suitable method for detection and removal of multiply charged peak clusters (deconvolution). Also, processing of MS/MS spectra with limited quality produced by low resolution MS instruments was facilitated by peak merging and spectra smoothing with a special median filter. As result of the background

removal procedures, the number of peaks in the spectra is reduced by one quarter in average. In this way, the quality of interpretable spectra increases which leads to more reliable interpretation results. On the other hand, non-interpretable spectra are recognized by a new, ingeniously simple algorithm searching for sets of peaks with substantial intensity having mass distances that correspond to amino acid residues (sequence ladders). In average, two thirds of the spectra are removed from further consideration. In total, the background removal (of non-interpretable spectra and of background peaks) results in saving of three quarters of the computation time that is necessary for MS/MS spectrum interpretation.

These algorithms and methods were implemented in a computer program called "MS Cleaner" which has become a standard step in proteomics studies on the Research Institute of Molecular Pathology Vienna, and is used before submitting MS/MS spectra to interpretation software. The program MS Cleaner outperforms any preexisting technique by an order of magnitude. The methods and algorithms have been submitted to the US Patent Office and partially published in an article of the journal "Proteomics" [33]. A second publication for the description of the sequence ladder criterion and the multiprocessor version of "MS Cleaner" is currently being finalized and will be submitted to the journal "Nature Methods".

This thesis includes the following chapters: After a description of the mass spectrometric workflow for the identification of proteins and a consideration of the scarce literature on MS/MS spectrum preprocessing, the experimental procedures used for generating the sample MS/MS data are reported. This data was essential for parameterizing the background removal procedures and for testing the performance of algorithms. These chapters are followed by the description of technical ideas that underlie the new background removal

procedures developed in this work. The respective algorithms are presented in pseudocode and with flow charts in chapter 5 and their implementation details are given in chapter 6. The performance of the new methods is evaluated in chapter 7.

Chapter 2

Previous Work on Treating Background in MS/MS Spectra

Background processing of raw MS/MS spectra from protein samples has not been in the center of interest among the community for a long time, partly due to limitations of measurement accuracy. For example, resolution of isotope clusters requires very precise instruments, which have become available on a broad scale only recently (for example, the Thermo Finnigan LCQ with close to ≈0.5 Da resolution or the newer LTQ with ≈0.3 Da resolution). Therefore, some spectrum interpretation algorithms foresee simplified exclusion rules for heavy ion peaks in their scoring or spectra pre-processing schemes [26]. Similarly, deconvolution of multiply charged peaks and deisotoping with procedures described in the literature [34, 35, 36, 37, 38, 39, 40, 41, 42] are possible only with very accurate data and resolved isotope clusters. The results are reliable only in cases of sufficiently large peptide fragments where an isotope peak cluster of the higher charge state is confirmed by respective clusters at the lowest charge state or when the distances between peaks in a cluster accurately match the expected mass differences.

Sometimes, it might be rather advisable to refrain from automatically interpreting very noisy MS/MS spectra instead of generating interpretations that are not justified by the data. The task of unselecting non-interpretable spectra is related to but different from the question of cleaning spectra from noise. Xu et al.[43] and Bern et al.[44] propose empirical criteria for unselecting bad spectra; i.e., spectra with only few significant peaks over a dense background. For these methods, the relatively high number of false-positively unselected (i.e., nevertheless interpretable) spectra remains a problem.

Previous work on raw protein MS/MS spectrum processing has not led to satisfying solutions and, therefore, many currently available MS/MS spectrum analysis packages largely ignore the presence of additional background signals. Most commercial spectrum interpretation software suites contain some noise reduction but the algorithms implemented are not publicly documented. At present, there is only one available program dedicated to spectral cleaning, the Mascot Distiller (see www.matrixscience.com), a commercial software package that optimizes peak location and intensities given the ideal isotopic distribution of elements contained in peptides. However, the algorithms used in this software are not published and the correctness of peak removal/inclusion has not been evaluated in transparent large-scale tests. In addition, low computation speed and run-time stability issues may create problems in practical lab work.

It should be emphasized that, given the incomplete understanding of the chemical process of fragmentation, no automated procedure will match the performance of the experienced eye and the intuition of a mass spectrometry specialist in the foreseeable future. Nevertheless, the number of mass spectra to be processed in proteomics laboratories is so large that there is no alternative to automated interpretation, maybe, augmented by manual inspection

of a few selected cases.

The considerations detailed above naturally lead to the following questions: Is it possible to detect repeated signals in MS/MS spectra like singly and multiply charged peaks which would disturb interpretation of MS/MS spectra if they were left in an MS/MS spectrum unmodified? Could we solve the general problem of mass spectrometry and algorithmically transform these signals into interpretable signals? Is it possible to reduce the amount of peaks in the spectrum by extracting only interpretable peaks? What is the smallest amount of all produced spectra to successfully identify a protein? What are the possibilities of finding non-informative spectra and how would the processing time and the result of protein identification benefit from detection of bad spectra?

Chapter 3

General Overview About Mass Spectrometry

Mass spectrometers can be divided into three fundamental parts, namely the ionization source, the analyzer, and the detector (Figure 3.1).

The sample under investigation has to be brought into the ionization source of the instrument. Once inside the ionization source, the sample molecules are ionized and the resulting ions are extracted into the analyzer region of the mass spectrometer. In the analyzer, they are separated according to their mass-to-charge ratios (m/z). The separated ions are detected and this signal is sent to a data system where the m/z ratios are stored together with their relative abundance for presentation in the format of an m/z spectrum.

The analyzer and detector of the mass spectrometer and, often, the ionization source, too, are maintained under high vacuum to give the ions a reasonable chance of traveling from one end of the instrument to the other without any hindrance from air molecules.

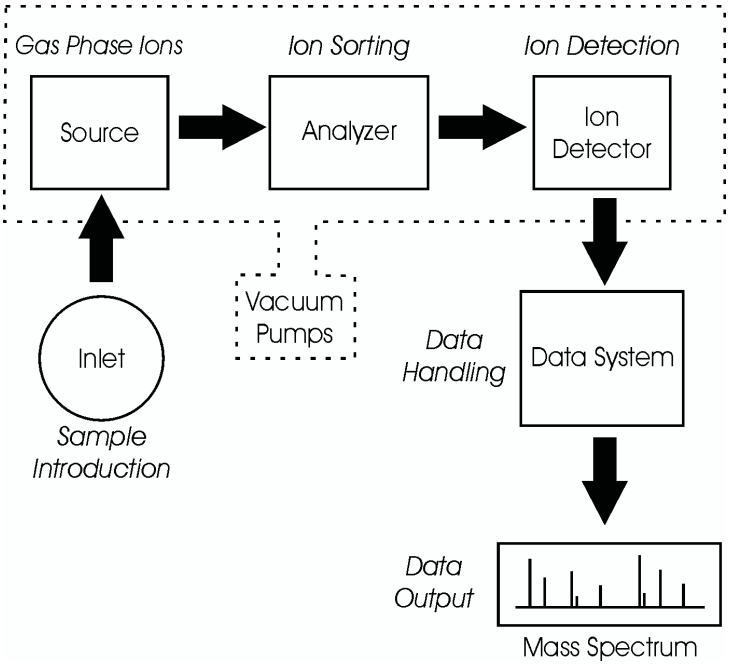


Figure 3.1: Simplified scheme of a mass spectrometer

3.1 Sample Introduction

The method of sample introduction to the ionization source often depends on the ionization method being used, as well as the type and complexity of the sample.

The sample can be inserted directly into the ionization source, or can undergo some type of chromatography prior to ionization. This latter method of sample introduction usually involves the mass spectrometer being coupled directly to a high pressure liquid chromatography (HPLC), gas chromatography (GC) or capillary electrophoresis (CE) separation column and, hence, the sample is separated into a series of components which enter the mass spectrometer sequentially for individual analysis[45].

3.2 Ionization Methods

The ionization method to be used should depend on the type of sample under investigation and the mass spectrometer available. The ionization methods used for the majority of biochemical analyses are Electrospray Ionization (ESI) and Matrix Assisted Laser Desorption Ionisation (MALDI)[45].

Electrospray Ionisation (ESI)[45] is one of the *Atmospheric Pressure Ionisation (API)* techniques and is well-suited to the analysis of polar molecules ranging from less than 100 Da to more than 1,000,000 Da in molecular weight.

During standard electrospray ionization [46], the sample is dissolved in a polar, volatile solvent and pumped through a narrow, stainless steel capillary (75 - 150 m i.d.) at a flow rate of between 1 L/min and 1 mL/min.

A high voltage of 3 or 4 kV is applied to the tip of the capillary (Figure 3.2), which is situated within the ionization source of the mass spectrometer, and as a consequence of this strong electric field, the sample emerging from the tip is dispersed into an aerosol of highly charged droplets, a process that is aided by a co-axially introduced nebulising gas flowing around the outside of the capillary. This gas, usually nitrogen, helps to direct the spray emerging from the capillary tip towards the mass spectrometer.

The charged droplets (Figure 3.3) diminish in size by solvent evaporation, assisted by a warm flow of nitrogen known as the drying gas which passes across the front of the ionization source. Eventually, charged sample ions, free from solvent, are released from the droplets, some of which pass through a sampling cone or orifice into an intermediate vacuum region and, from there, through a small aperture into the analyzer of the mass spectrometer, which is held under high vacuum. The lens voltages are optimized individually for each sample.

Nanospray ionization[47] is a low flow rate version of electrospray ioniza-

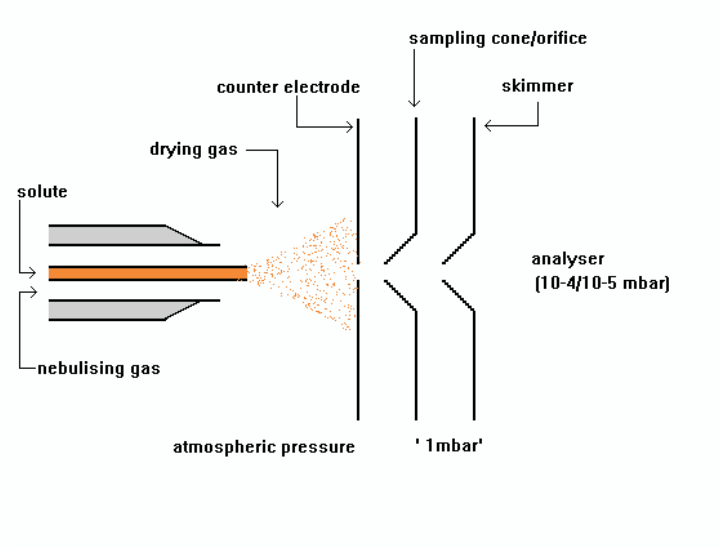


Figure 3.2: Standard electrospray ionization source

tion. A small volume (1-4 L) of the sample dissolved in a suitable volatile solvent, at a concentration of ca. 1 - 10 pmol/L, is transferred into a miniature sample vial. A reasonably high voltage (ca. 700 - 2000 V) is applied to the specially manufactured gold-plated vial resulting in sample ionization and spraying (Figure 3.4).

Desolvation is followed by ion extraction through the sampling cone, which is situated at 90 ° to the original flow of solute and solvent and, then, through the extraction cone (another 90 ° turn) into the analyzer for separation and analysis of the ions according to their m/z ratios, as with standard ESI-MS. The two right-angled bends in the ionization source have led to its name of Z-spray.

The flow rate of solute and solvent using this procedure is very low, 30 - 1000 nL/min. Thus, not only is far less sample consumed than with the standard electrospray ionization technique, but also a small volume of sample lasts for several minutes, enabling multiple experiments to be performed.

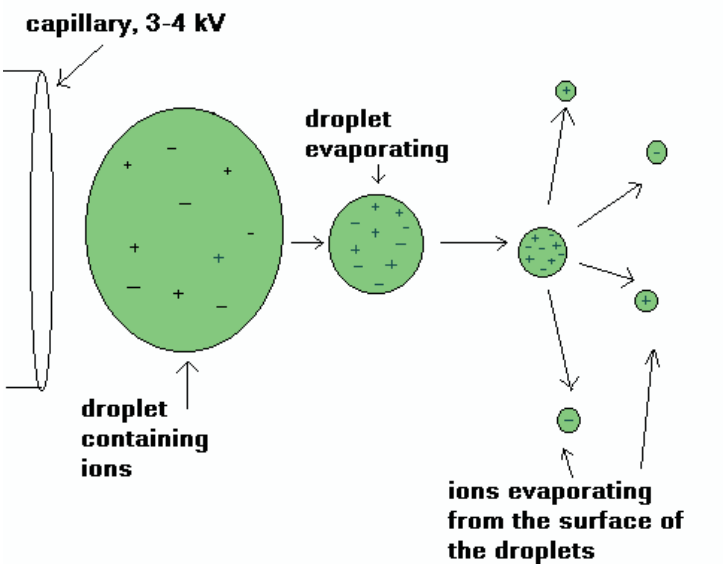


Figure 3.3: The electrospray ionization process

A common application of this technique is for a protein digest mixture to be analyzed to generate a list of molecular weights for the components present and, then, each component to be analyzed further by tandem mass spectrometric (MS-MS) amino acid sequencing techniques.

Matrix Assisted Laser Desorption Ionization (MALDI)[48] deals well with thermo-labile, non-volatile organic compounds especially those of high molecular weight and is used successfully in biochemical areas for the analysis of proteins, peptides, glycoproteins, oligosaccharides, and oligonucleotides. It is relatively straightforward to use and reasonably tolerant to buffers and other additives.

The mass accuracy depends on the type and performance of the analyzer of the mass spectrometer, but most modern instruments should be capable of measuring masses to within 0.01% of the molecular weight of the sample, at least up to ca. 40,000 Da.

MALDI is based on the bombardment of sample molecules with a laser

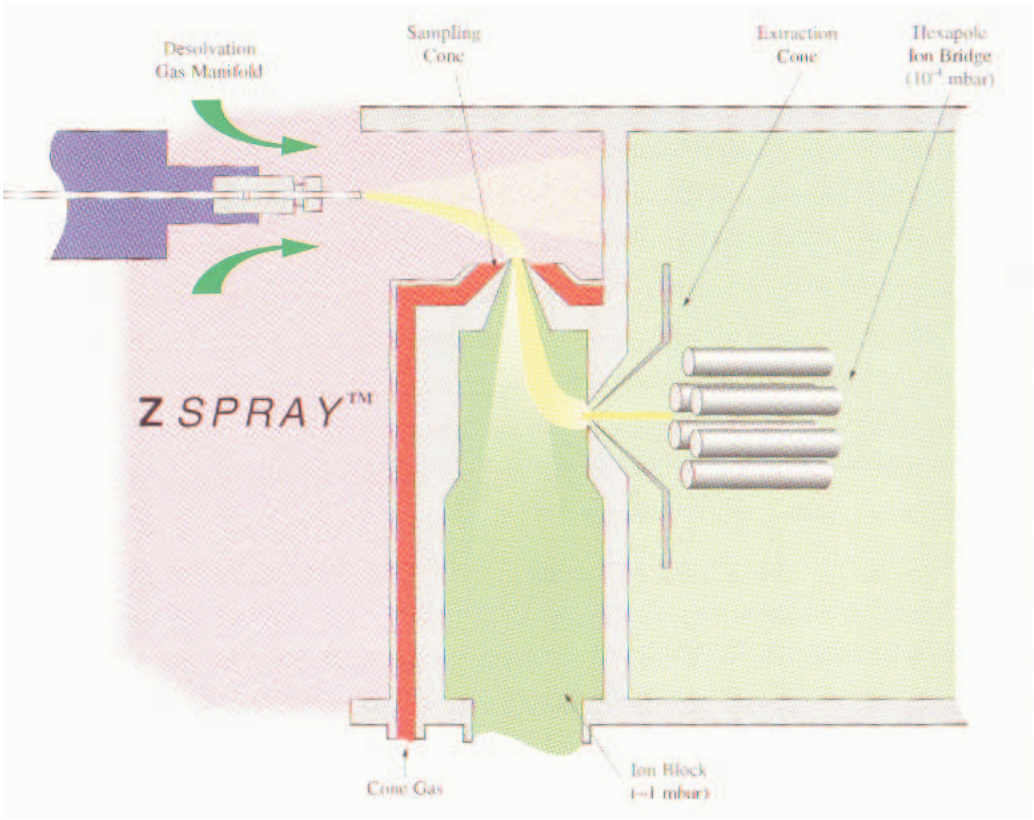


Figure 3.4: Nanospray ionisation process using a Z-Spray ionisation source (Q-TOF)

light to bring about sample ionization (Figure 3.5). The sample is pre-mixed with a highly absorbing matrix compound for the most consistent and reliable results, and a low concentration of sample to matrix works best. The matrix transforms the laser energy into excitation energy for the sample, which leads to desorption of analyte and matrix ions from the surface of the mixture. In this way, energy transfer is efficient and also the analyte molecules are spared excessive direct energy that may otherwise cause decomposition. Most commercially available MALDI mass spectrometers now have a pulsed nitrogen laser of wavelength 337 nm.

The sample to be analyzed is dissolved in an appropriate volatile solvent,

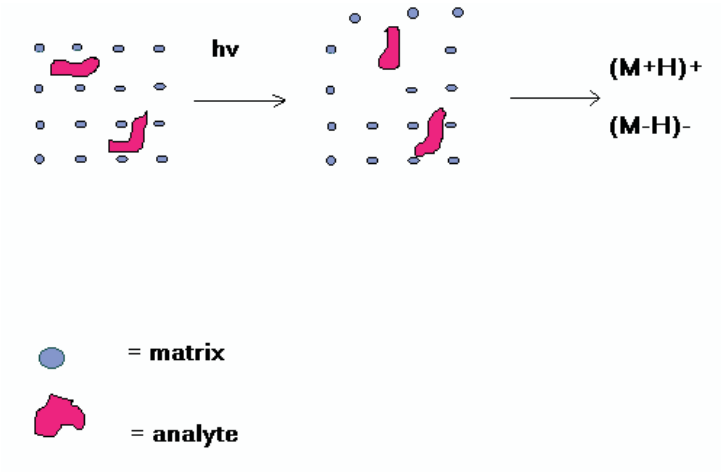


Figure 3.5: Matrix Assisted Laser Desorption Ionization (MALDI)

usually with a trace of trifluoroacetic acid if positive ionization is being used, at a concentration of ca. 10 pmol/L and an aliquot (1-2 μ L) of this removed and mixed with an equal volume of a solution containing a vast excess of a matrix.

A range of compounds is suitable for use as matrices: sinapinic acid is a common one for protein analysis while α -cyano-4-hydroxycinnamic acid is often used for peptide analysis. An aliquot (1-2 μ L) of the final solution is applied to the sample target which is allowed to dry prior to insertion into the high vacuum of the mass spectrometer. The laser is fired, the energy arriving at the sample/matrix surface is optimized, and data is accumulated until a m/z spectrum of reasonable intensity has been amassed.

The time-of-flight analyzer separates ions according to their mass(m)-to-charge(z) (m/z) ratios by measuring the time it takes for ions to travel through a field free region known as the flight, or drift, tube. The heavier ions are slower than the lighter ones 3.6.

The m/z scale of the mass spectrometer is calibrated with a known sample that can either be analyzed independently (external calibration) or pre-mixed

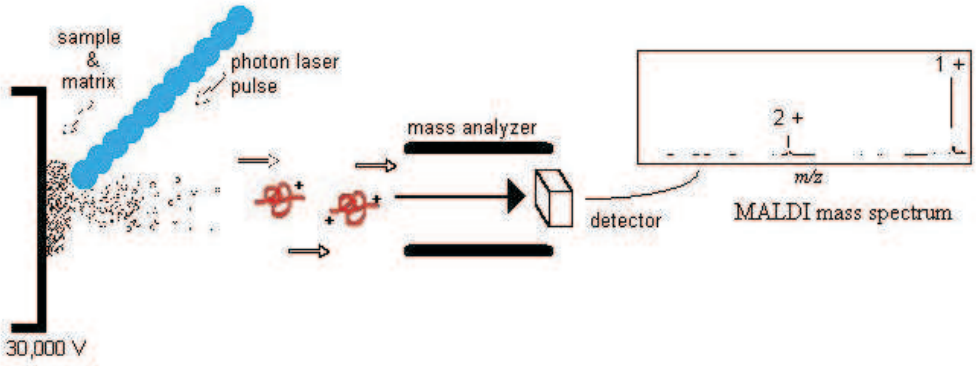


Figure 3.6: Simplified scheme of MALDI-TOF mass spectrometry

with the sample and matrix (internal calibration).

MALDI is a "soft" ionization method. So, it results predominantly in the generation of singly charged molecular-related ions regardless of the molecular weight. Hence, the spectra are relatively easy to interpret. Fragmentation of the sample ions does not usually occur.

In the positive ionization mode, the protonated molecular ions (M+H+) are usually the dominant species, although they can be accompanied by salt adducts, a trace of the doubly charged molecular ion at approximately half the m/z value, and/or a trace of a dimeric species at approximately twice the m/z value. Positive ionization is used in general for protein and peptide analyses.

In the negative ionization mode, the deprotonated molecular ions (M-H-) are usually the most abundant species, accompanied by some salt adducts and possibly traces of dimeric or doubly charged materials. Negative ionization can be used for the analysis of oligonucleotides and oligosaccharides.

3.3 Analysis and Separation of Sample Ions

The main function of the mass analyzer is to separate, or resolve, the ions formed in the ionization source of the mass spectrometer according to their mass-to-charge (m/z) ratios. There are a number of mass analyzers currently available, the better known of which include quadrupoles, time-of-flight (TOF) analyzers, magnetic sectors, and both Fourier transform and quadrupole ion traps.

These mass analyzers have different features, including the m/z range that can be covered, the mass accuracy, and the achievable resolution. The compatibility of different analyzers with different ionization methods varies. For example, all of the analyzers listed above can be used in conjunction with electrospray ionization, whereas MALDI is not usually coupled to a quadrupole analyzer.

The *single sector magnetic mass analyzer* uses only a magnetic field to separate ions with different mass-to-charge ratios. The ions entering the mass analyzer are initially accelerated using an electric field and only ions with a certain charge are passed through. Then, these ions enter a magnetic field. Charged ions tend to move in a circular trajectory in a magnetic field depending on their mass and, thus, reach the ion detector at different locations. The *double sector mass analyzer* uses an additional electric field to filter ions such that only ions with a certain kinetic energy are passed through to the magentic sector. The ions are then separated according to their mass in the magetic sector as before.

The quadrupoles in the quadrupole mass analyzer (Figure 3.7) are 4 parallel rods that are controlled by DC voltage and also an RF potential. Ions with specific mass-to-charge ratios can be separated by controlling the RF potential. Quadrupole analyzers are characterized by their insensitivity to

poor vacuum, low cost and ability to measure high mass-to-charge ratios.

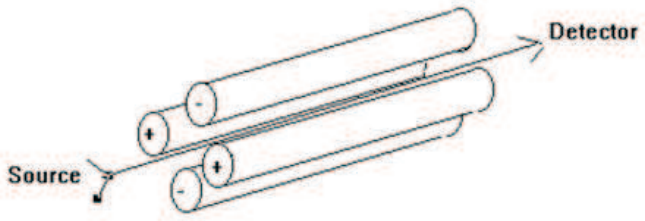


Figure 3.7: Scheme of quadrupole analyzer

The *quadrupole ion trap mass analyzer* (Figure 3.8) is similar to the *quadrupole analyzer*. Here, the ions of interest with a specific mass-to-charge ratio are trapped inside a radio frequency quadrupole field. Ions can be ejected from the ion trap by changing the RF potential. So by changing the RF potential, one can eject ions with different mass-to-charge ratios from ion trap sequentially and each species can be further analyzed separately without performing different experiments.

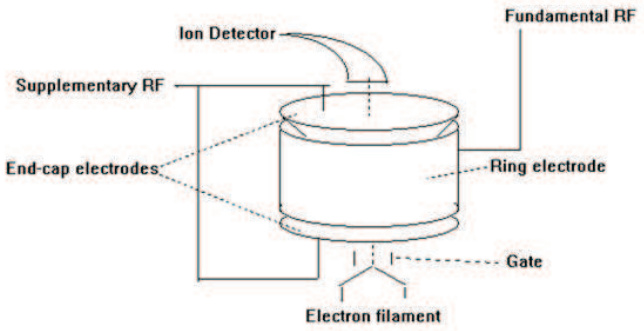


Figure 3.8: Scheme of quadrupole ion trap mass analyzer

In a *time-of-flight mass analyzer* (Figure 3.9) the different ions are accelerated down a cylinder towards the ion detector with the same energy. Since different ions might have different masses, the ions will reach the detector

at different times with smaller ions reaching the detector before the larger ions. The mass of the ions is determined from the time of arrival, which is a function of mass, charge and time of travel of the ion.

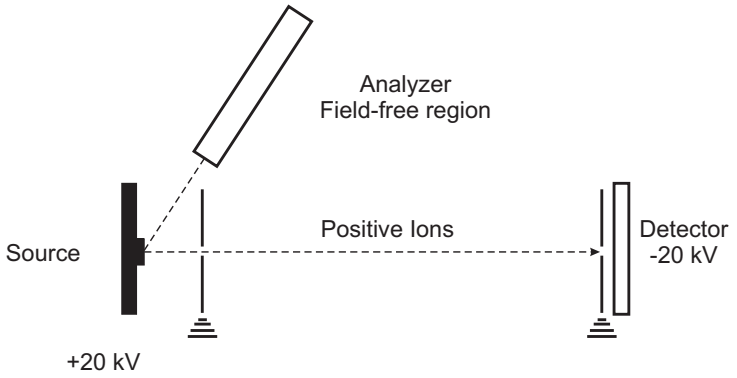


Figure 3.9: Scheme of Time-Of-Flight mass analyzer

3.4 Detection and Recording of Sample Ions

The detector monitors the ion current, amplifies it and the signal is transmitted to the data system where it is recorded in the form of mass spectra. The m/z values of the ions are plotted against their intensities to show the number of components in the sample, the molecular weight of each component, and the relative abundance of the various components in the sample.

The type of detector is supplied to suit the type of analyzer; the more common ones are the photomultiplier, the electron multiplier and the microchannel plate detectors.

3.5 Tandem Mass Spectrometry

Tandem mass spectrometry (MS-MS) is used to produce primary structural information about a compound by fragmenting specific sample ions inside the

mass spectrometer and identifying the resulting fragment ions. This information can then be pieced together to generate structural information regarding the intact molecule. Tandem mass spectrometry also enables specific compounds to be detected in complex mixtures on account of their specific and characteristic fragmentation patterns.

A tandem mass spectrometer is a mass spectrometer that has more than one analyzer, in practice usually two. The two analyzers are separated by a collision cell into which an inert gas (e.g. argon, xenon) is admitted to collide with the selected sample ions and bring about their fragmentation. The analyzers can be of the same or of different types, the most common combinations being:

- quadrupole - quadrupole
- magnetic sector - quadrupole
- magnetic sector - magnetic sector
- quadrupole - time-of-flight.

The Q-Tof mass spectrometer is a *quadrupole-time-of-flight* tandem mass spectrometer. Fragmentation experiments can also be performed on certain single analyzer mass spectrometers such as *ion trap and time-of-flight* instruments, the latter type using a post-source decay experiment to effect the fragmentation of sample ions.

The basic modes of data acquisition for tandem mass spectrometry experiments are as follows:

- *Product or daughter ion scanning*: The first analyzer is used to select user-specified sample ions arising from a particular component, usually

the molecular-related (i.e. (M+H)+ or (M-H)-) ions. These chosen ions pass into the collision cell, are bombarded by the gas molecules which cause fragment ions to be formed, and these fragment ions (i.e., separated according to their mass to charge ratios) are analyzed by the second analyzer. All fragment ions arise directly from the precursor ions specified in the experiment and, thus, produce a fingerprint pattern specific to the compound under investigation. This type of experiment is particularly useful for providing structural information concerning small organic molecules and for generating peptide sequence information.

- *Precursor or parent ion scanning*: The first analyzer allows the transmission of all sample ions, whilst the second analyzer is set to monitor specific fragment ions, which are generated by bombardment of the sample ions with the collision gas in the collision cell. This type of experiment is particularly useful for monitoring groups of compounds contained within a mixture which fragment to produce common fragment ions, e.g. glycosylated peptides in a tryptic digest mixture, aliphatic hydrocarbons in an oil sample, or glucuronide conjugates in urine.

- *Constant neutral loss scanning*: This involves both analyzers scanning, or collecting data, across the whole m/z range, but the two are off-set so that the second analyzer allows only those ions which differ by a certain number of mass units (equivalent to a neutral fragment) from the ions transmitted through the first analyzer. E.g., this type of experiment could be used to monitor all of the carboxylic acids in a mixture. Carboxylic acids tend to fragment by losing a (neutral) molecule of carbon dioxide, CO2, which is equivalent to a loss of 44 Da or atomic

mass units. All ions pass through the first analyzer into the collision cell. The ions detected from the collision cell are those from which 44 Da have been lost.

- *Selected/multiple reaction monitoring*: Both of the analyzers are static in this case as user-selected specific ions are transmitted through the first analyzer and user-selected specific fragments arising from these ions are measured by the second analyzer. The compound under scrutiny must be known and have been well-characterized before this type of experiment is undertaken. This methodology is used to confirm unambiguously the presence of a compound in a matrix, e.g. drug testing with blood or urine samples. It is not only a highly specific method but also has very high sensitivity.

3.6 Peptide Analysis Using Mass Spectrometry

Figure 3.10 shows the process of obtaining a mass spectrum for a sample containing the peptide of interest whose identity is to be determined. The peptide to be analyzed is first separated from the mixture of peptides and purified (using GC, HPLC, etc.) before it is introduced into the mass spectrometer. The peptide is then subject to mass spectrometry and its mass spectrum is obtained. The mass-to-charge ratio and intensity of the ions in the mass spectrum can be used to identify the unknown peptide[45].

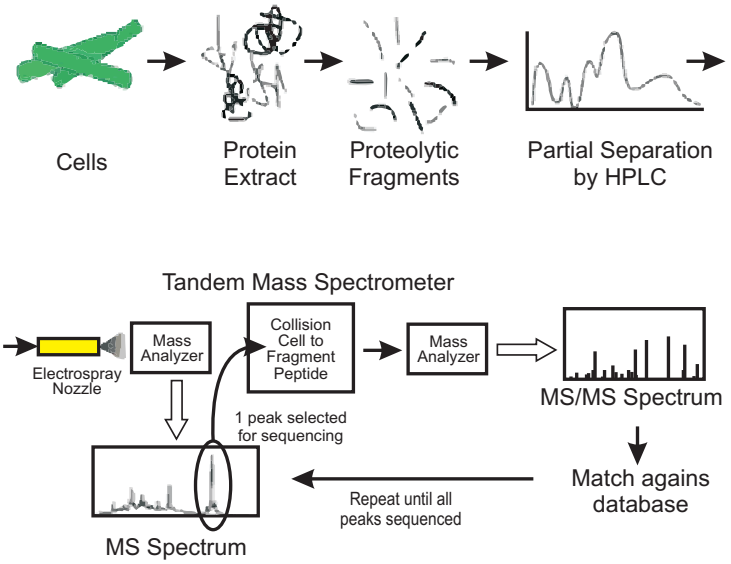


Figure 3.10: Peptide analysis using mass spectrometry

3.7 Peptide Sequencing by Tandem Mass Spectrometry

The most common usage of MS-MS in biochemical areas is the product or daughter ion scanning experiment which is particularly successful for peptide and nucleotide sequencing.

Peptides decay in a reasonably well-documented manner [49, 50]. The protonated molecules fragment along the peptide backbone (Figure 3.11) and also show some side-chain fragmentation [51].

There are three different types of bonds that can fragment along the amino acid backbone: the NH-CR, CR-CO, and CO-NH bonds. Each bond breakage gives rise to two species, one neutral and the other one charged, and only the charged species is monitored by the mass spectrometer. The charge can stay on either of the two fragments depending on the chemistry and relative proton affinity of the two species. Hence there are six possible

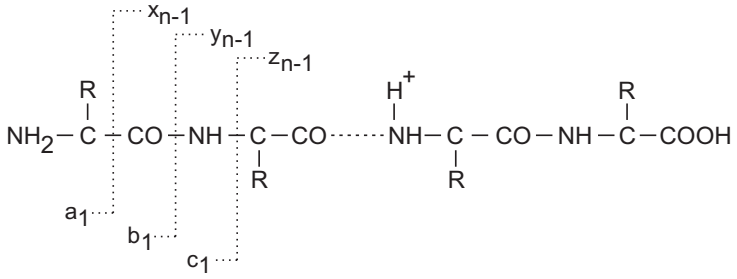


Figure 3.11: Peptide sequencing by tandem mass spectrometry - backbone cleavages

fragment ions for each amino acid residue and these are labeled as in the diagram, with the a, b, and cïons having the charge retained on the N-terminal fragment, and the x, y; and z ions having the charge retained on the C-terminal fragment. The most common cleavage sites are at the CO-NH bonds which give rise to the b and (or) the yïons.

The extent of side-chain fragmentation detected depends on the type of analyzers used in the mass spectrometer. A magnetic sector - magnetic sector instrument will give rise to high energy collisions resulting in many different types of side-chain cleavages. Quadrupole - quadrupole and quadrupole - time-of-flight mass spectrometers generate low energy fragmentations with fewer types of side-chain fragmentations.

Immonium ions (H_2N^+=CHR) appear in the very low m/z range of the MS-MS spectrum. Each amino acid residue leads to a diagnostic immonium ion, with the exception of the two pairs leucine (L) and iso-leucine (I), and lysine (K) and glutamine (Q), which produce immonium ions with the same m/z ratio, i.e. m/z 86 for I and L, m/z 101 for K and Q. The immonium ions are useful for detecting and confirming many of the amino acid residues in a peptide, although no information regarding the position of these amino acid residues in the peptide sequence can be ascertained from the immonium ions.

A protein identification study would typically proceed as follows:

a) The protein is digested with a suitable enzyme. Trypsin is useful for mass spectrometric studies because each proteolytic fragment contains a basic arginine (R) or lysine (K) amino acid residue and, thus, is eminently suitable for positive ionization mass spectrometric analysis. The digest mixture is analyzed - without prior separation or clean-up - by mass spectrometry to produce a rather complex spectrum from which the molecular weights of all of the proteolytic fragments can be read. This spectrum, with its molecular weight information, is called a peptide map (peptide fingerprint). (If the protein already exists in a database, then the peptide map is often sufficient to confirm the identity of the protein.) For these experiments, the Q-Tof mass spectrometer would be operated in the "MS" mode (Figure 3.12) , whereby the sample is sprayed and ionized from the nanospray needle and the ions pass through the sampling cone, skimmer lenses, RF hexapole focusing system, and the first (quadrupole) analyzer. The quadrupole in this instance is not used as an analyzer, merely as a lens to focus the ion beam into the second (time-of-flight) analyzer which separates the ions according to their mass-to-charge ratio.

b) With the digest mixture still spraying into the mass spectrometer, the Q-Tof mass spectrometer is switched into "MS-MS" mode (Figure 3.13). The protonated molecular ions of each of the digest fragments can be independently selected and transmitted through the quadrupole analyzer, which is now used as an analyzer to transmit solely the ions of interest into the collision cell which lies in-between the first and second analyzers.

An inert gas such as argon is introduced into the collision cell and the sample ions are bombarded by the collision gas molecules which cause them to fragment. The optimum collision cell conditions vary from peptide to peptide

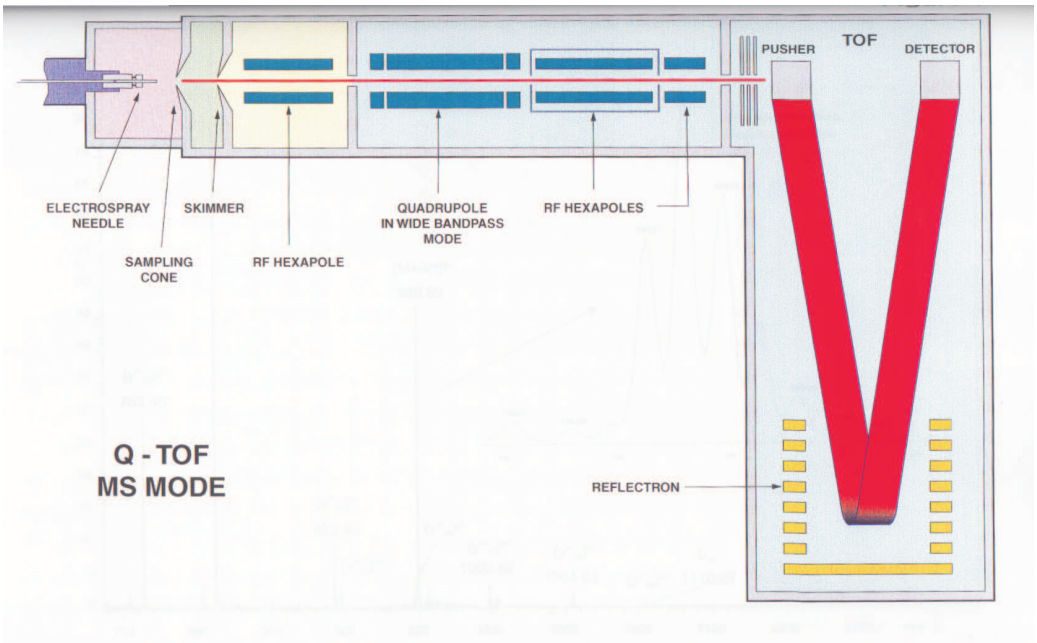


Figure 3.12: Q-TOF operating in MS-MS mode

and must be optimized for each one. The fragment (or daughter or product) ions are then analyzed by the second (time-of-flight) analyzer. In this way, an MS-MS spectrum is produced showing all the fragment ions that arise directly from the chosen parent or precursor ions for a given peptide component.

An MS-MS daughter (or fragment, or product) ion spectrum is produced for each of the components identified in the proteolytic digest. Varying amounts of sequence information can be obtained from each fragmentation spectrum and the spectra need to be interpreted carefully. Some of the processing can be automated but, in general, the processing and interpretation of spectra will take longer than the data acquisition if accurate and reliable results are to be generated.

The proteomics procedure usually involves excising individual spots from a 2-D gel and independently enzymatically digesting the protein(s) contained

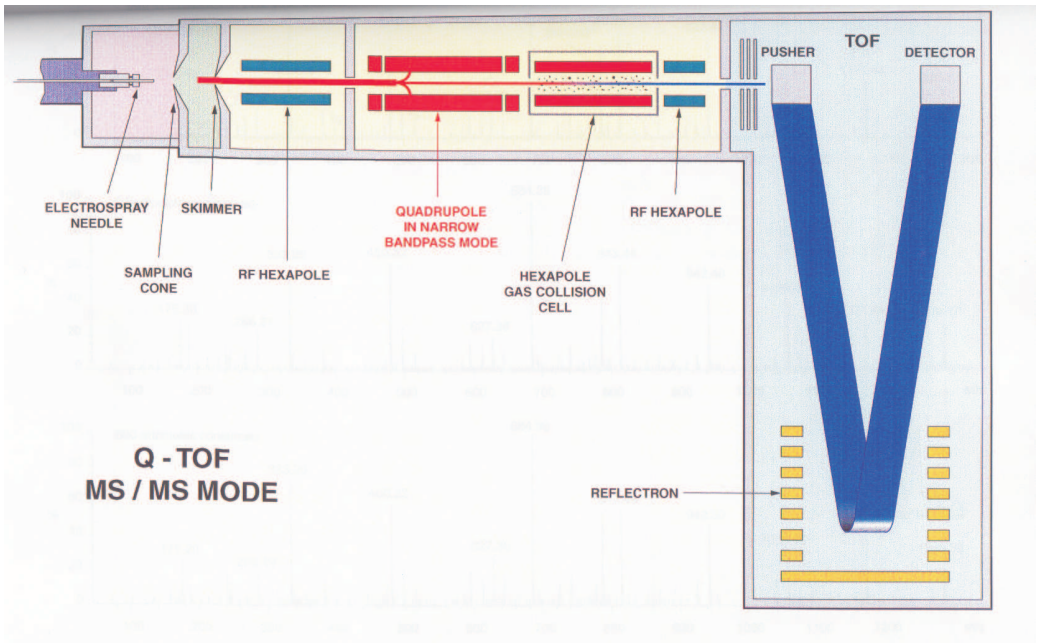


Figure 3.13: Q-TOF operating in MS mode

within each spot, before analyzing the digest mixture by mass spectrometer in the manner outlined above.

Chapter 4

Experimental Procedure for Obtaining MS/MS Spectra

To develop robust algorithms and methods for spectra pre-processing and cleaning it is necessary to work with real data. For this purpose, cell extract proteins obtained from IMP laboratories as well as commercially acquired proteins were used for MS analysis.

4.1 Sample Preparation

Cell extract proteins obtained from IMP laboratories are prepared with following steps: 200 g of purified anti-human Smc2 rabbit polyclonal antibody [52], crosslinked to Affi-Gel Protein A beads (100 μ L bed-volume, Bio-Rad), was used to immunoprecipitate the condensin complexes from 10 mg of clarified interphase HeLa cell extract. Following extensive washing, immunoprecipitated protein complexes were acid-eluted from the beads, and 10% of the total eluate was analysed by SDS-PAGE and silver staining. After reduction and acetylation of cysteine residues using dithiothreitol and iodoacetamide,

respectively, the condensin sample was proteolytically digested using Trypsin Gold (Promega), and the digestion stopped with tetrafluoroacetic acid.

Commercially acquired proteins are: α-amylase, amylogucosidase, apo-transferrin, β-galactidase, carbonic anhydrase, catalase, phosphorylase B, glutamic dehydrogenase, glutathione transferase, immunoglobulin γ, lactic dehydrogenase, lactoperoxidase, myoglobin.

4.2 Mass Spectrometry

Tryptic peptides from condensin samples were separated by nano-HPLC[53] on an UltiMate HPLC system and PepMap C^{18} column (LC Packings, Amsterdam, The Netherlands), with a gradient of 5-75% acetonitrile, in 0.1% formic acid[54, 55]. Eluting peptides were introduced by electrospray ionisation (ESI) into an LTQ linear ion trap mass spectrometer (Thermo Finnigan), where full MS and MS/MS spectra were recorded. In another experiment, a mixture of tryptic peptides from standard, commercially acquired bovine serum albumin (BSA), yeast alcohol dehydrogenase (ADH) or human transferrin (TRF) were used for system optimization and testing. 100 fmol of each protein were injected into a nanoHPLC device (LC Packings, Amsterdam, The Netherlands) and MS/MS spectra were acquired using a 3D ion trap mass spectrometer, model DecaXP (Thermo Finnigan).

Commercially acquired proteins were used, each in two preparations. For chromatography, a UltiMate Plus Nano-LC system. LC-Packings - A Dionex Co was applied. The sample was loaded (loading solvent: water; 0.1% TFA) for 10 min onto a reversed phase trap column (which is not online with the separation column; description: PepMap C18, 300 m ID x 5mm length, 3 m particle size, 100 pore size, LC Packings - A Dionex Co.) at a flow

rate of 20 l/min and washed free of ion pairing agents and other impurities. The gradient described starts at 10 min (the trap column is switched online with the separation column, mobile phase: 95% water, 5% acetonitrile, 0.1% FA, flow rate 0.275 l/min) and continues for 50 min. After applying a high organic wash step (95% mobile phase with 20% water, 80% acetonitrile, 0.1% FA), the trap column is switched back to offline mode and equilibrated with the loading mobile phase. The mass spectrometric data are recorded only for the time both columns are online. The mass spectra were recorded with a Thermo Finnigan LTQ (positive nano-ESI mode, ionizing spray voltage: 1.5 kV, enhanced mass-spec full-scan range: 220 - 2000 amu).

4.3 File Processing

The MS/MS output in the Xcalibur raw-file was converted into dta-files using BioWorks (by thermo.com). Dta-files are text files with following format: The first row contains the mass and the charge state of the precursor ion from which the MS/MS spectrum was generated. All following rows contain m/z values in the first column and the intensity in the second column. Single dta-files were used to examine possibilities of spectra cleaning and pre-processing.

In order to check benefits of applying different algorithms and methods for spectra cleaning and preprocessing, the MS/MS spectra interpretation software called Mascot[30] was used. The respective dta-files were merged to generate a single mgf-file (Mascot generic format) using the merge.pl program from Matrix Science (www.matrixscience.com). This original mgf-file, which is a collection of dta files, was processed by Mascot. Improved recognition and protein identification by Mascot was considered as main criterion for accepting new developments of algorithms and methods.

Chapter 5

Algorithms for Cleaning and Pre-Processing of MS/MS Spectra

5.1 Introduction

For a given raw (but centroided) peptide MS/MS spectrum in dta format, five independent procedures were developed:

(i) for detection of multiply charged peaks (the algorithm "Deconvolute Spectrum"),

(ii) for removal of latent periodic noise including deisotoping (the algorithm "Deisotope Spectrum"),

(iii) for removal of high-frequency random noise (the algorithm "Remove Random Noise"),

(iv) for detection of non-interpretable spectra using information content

from PS (the algorithm "Deisotope Spectrum"), and

(v) for detection of non-interpretable spectra using sequence ladder test (the algorithm "Check Sequence Ladder").

Albeit comprehending the exact mechanism of the genesis of background peaks would allow the construction of an algorithm for their removal, this knowledge is not available and more phenomenological approaches appear necessary. The analogy with electrical signal processing is one possibility; i.e., the series of peaks in the mass spectrum can be considered as a signal compounded with noise after transfer via an information channel, from which the original signal has to be recovered.

The simplified cleaning and pre-processing procedure is shown in Figure 5.1.

5.2 The Algorithm "Check Sequence Ladder"

In this section, an idea from the beginning of mass spectrometry of proteins was used. Originally, experts tried to find amino acid sequence ladders among the high-intensity peaks. The computational costs are low to check in a MS/MS spectrum whether small ladders of predefined length do occur at all among the top fraction of most intense peaks. It is reasonable to suggest that the spectrum is probably not interpretable into a peptide sequence with statistical significance if no peptide sequence is matched by this criterion. Considerable amounts of MS/MS spectra origin from some non-peptide compounds present in the probe. Such compounds are mostly preparation artefacts, non-peptide polymers and other contaminants. On the other hand, peptide MS/MS spectra contain peaks with m/z values which differ from each other by amino-acid masses. In this work, the sequence ladder test had been

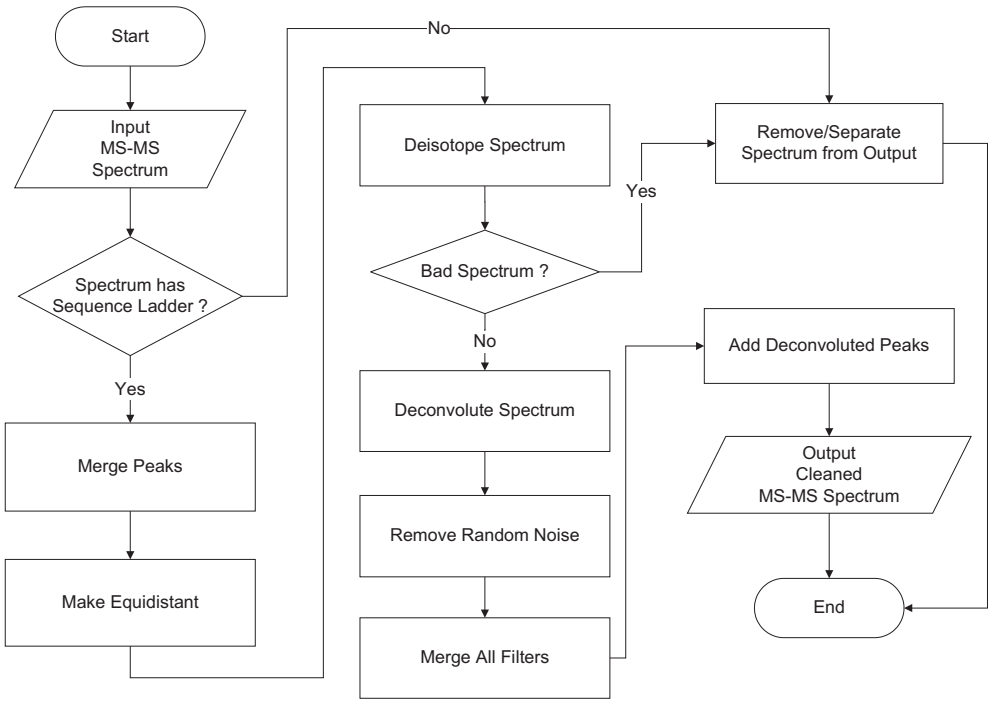


Figure 5.1: Simplified schema of spectra cleaning and pre-processing developed in this work

used to separate peptide MS/MS from other spectra. One might think that such a constraint is not generally applicable considering that spectra can contain multiply charged fragment ions. In practice, not all peaks are multiply charged and a relatively short (3-4 amino acid residues) ladder of singly charged peaks is found also in spectra that contain multiply charged peaks. From the result chapter, it can bee seen that such a simplification does not impact negatively the cleaning procedure.

The sequence ladder test algorithm checks an MS/MS spectrum for sequences of peaks that could describe an amino acid sequence. The output depends on input parameters and mass spectrometer resolution. It is a fast method to separate non-peptide spectra from the set of all spectra.

For the purpose of this algorithm, a "peak" is a tuple $< x, y >$ where x

is the peak's mass-to-charge (m/z) value and y its intensity. Given a peak p, i_p and m_p respectively represent the intensity and mass-to-charge ratio of p.

Check Sequence Ladder

Require:

	S	Set of peaks
	A	Set of amino acid masses
	msl	Minimum sequence ladder length to be found
	mt	Mass tolerance
	ip	percentage of highest intensity peaks to be included in search
	$haam$	Highest amino acid mass = 186.1 Da

1: $sll \leftarrow 0$
2: Find $S' \subset S$ such that $|S'| = |S| \cdot ip \quad \wedge \quad \forall\, x \in S' : \quad \nexists\, y \in S \setminus S'$ such that $i_y > i_x$
3: $k \leftarrow 0$
4: $j \leftarrow k + 1$
5: **for all** peaks $p(i) \in S'$ **do**
6: $\quad \Delta m \leftarrow |m_{p(j)} - m_{p(k)}|$
7: $\quad ba \leftarrow false$
8: $\quad$ **while** $\Delta m < haam \wedge ba = false$ **do**
9: $\quad\quad$ **if** $\exists a \in S \quad$ such that $\quad |a - \Delta m| < mt$ **then**
10: $\quad\quad\quad k \leftarrow j$
11: $\quad\quad\quad sll \leftarrow sll + 1$
12: $\quad\quad\quad$ **if** $sll \geqq msl$ **then**
13: $\quad\quad\quad\quad$ **return** Sequence ladder found
14: $\quad\quad\quad$ **end if**
15: $\quad\quad\quad ba \leftarrow true$

16: **end if**
17: **end while**
18: $j \leftarrow j + 1$
19: **end for**
20: **return** Sequence ladder not found

The first step is to extract an ordered subset (S') of required size containing the most intense peaks. A sequence starts with each peak $p(i) \in S'$ if there is a peak $p(j) \in S'$ such that distance between them is equal to a resudial mass of an amino acid. The sequence is extended until the required length is found or until all peaks in S' have been checked. If an amino acid sequence of required length could not be founded, the algorithm declares the spectrum as a non-interpretable spectrum.

The time complexity of the algorithm is $O(N^2)$ where N is the number of peaks in the spectrum (several hundreds). The quadratic complexity is the worst case and can be reduced if only neighbour peaks are checked if their m/z difference is equal to a mass of an amino acid. The neighbourhood width corresponds to the highest amino acid mass. The average case would then be $O(N \cdot M)$ where M is the number of peaks in neighbourhood and in real spectra it takes values $1 < M << 100$.

5.3 The Algorithm "Merge Peaks"

This algorithm merges a small intensity peak to a higher intensity neighbor peak if the m/z distance between them is under some certain value. Although the algorithm can be applied as a standalone noise removal procedure of minor peaks, it was developed to be used as a first step before a spectrum is deconvoluted with the algorithm "Deconvolute Spectrum" (section 5.7).

Minor peaks found within isotope peak clusters are artifacts that can arise from random noise or from the transformation of the continuous MS/MS spectrum into the centroid form as a discrete signal. The interfering peaks between main isotope cluster peaks have to be merged with the closest main heavy isotope peak in the cluster. Figure 5.2 and Figure 5.3 (enlarged) depict this problem.

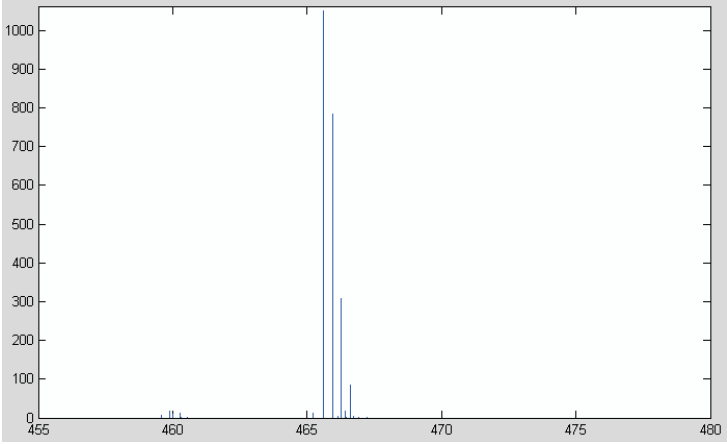


Figure 5.2: An m/z range showing small peaks between heavy isotope peaks

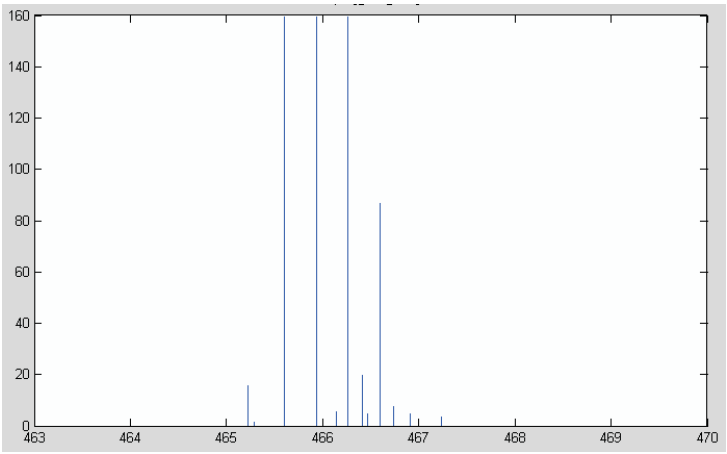


Figure 5.3: Enlarged view of an m/z range showing small peaks between heavy isotope peaks

For the peak-merging algorithm, a weighted directed acyclic graph $G(V, E)$

is constructed (Figure 5.4).

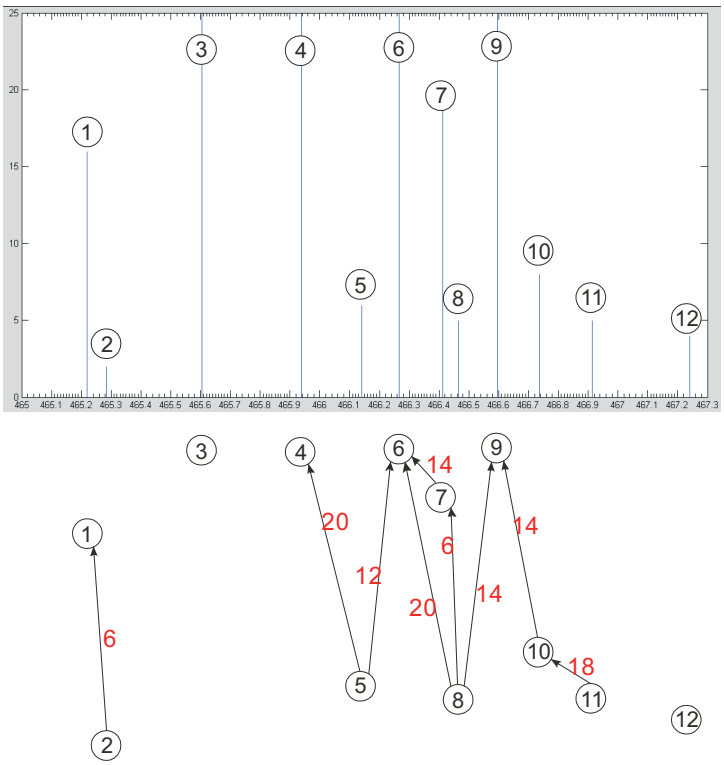


Figure 5.4: An m/z range converted into weighted directed acyclic graph

The set of vertices (V) is the set of all mass-over-charge values in the window. A directed edge $e_{i,j} \in E$ is added between two vertices $v_i, v_j \in V$ if the distance d (in Figure 5.4 depicted with red color, multiplied by 100) between peaks v_i, v_j is less than a certain value. The direction of the edge is defined to be from v_i to v_j if $Intensity(v_i) < Intensity(v_j)$. The weight w_i of an edge $e_{i,j}$ is defined as distance between two vertices v_i and v_j (in 0.01 Da units).

The algorithm "Deconvolute" requires no interfering minor peaks in the isotope peak cluster. The algorithm "Merge Peaks" creates a new graph $G'(V', E') \subset G(V, E)$ with $V' \subset V$ and $E' = \emptyset$ (Figure 5.5).

The removal of peaks from an MS/MS spectrum is generally not advis-

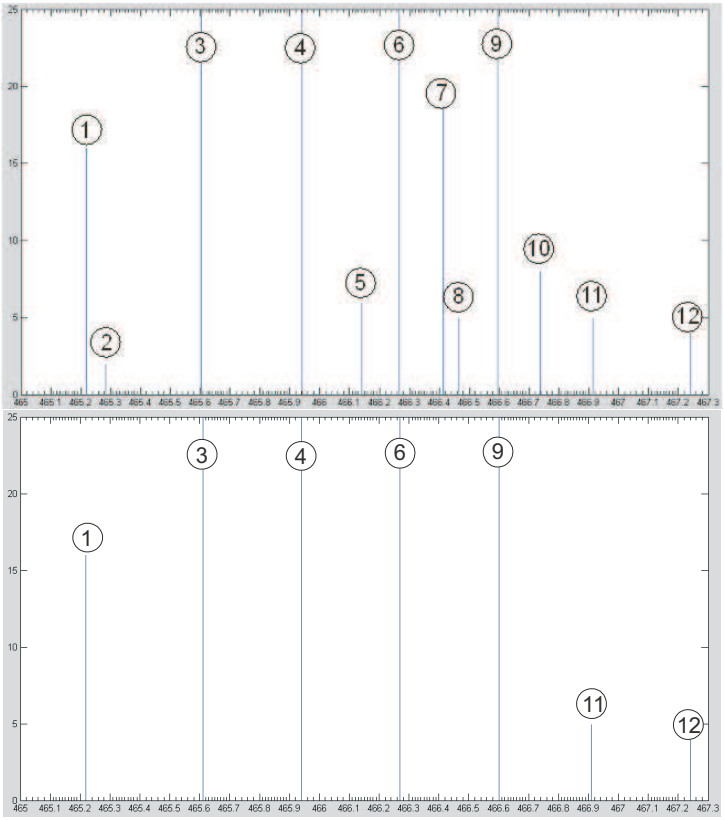


Figure 5.5: Enlarged m/z range before and after merging disturbing peaks

able, because some low intensity peaks are still peptide fragmentation peaks. The intensity of such peaks is low if the fragmentation along that particular peptide bond does not occur that often. For this reason $|V'|$ has to be as large as possible. This can be achieved if the sum of weights $w'_{i,j}$ of all removed edges $e'_{i,j}$ is as small as possible.

If a vertex v_i giving origin of the edge $e_{i,j}$ is actively removed from the graph (and its intensity is added to the vertex v_j), then edges to other vertices can also vanish.

For the purpose of this algorithm, three properties have been defined for each vertex. Given $v \in G(V, E)$, m_v and i_v represent the m/z value and original index of the corresponding peak in an MS/MS spectrum. $OutE_v \subset E$

is defined as set of all arcs with vertex v as tail (all out-going arcs from v).

Merge Peaks

Require: S Set of peaks

ld the lowest allowed distance between peaks

1: Create an emtpy weighted directed acyclic graph $G(V, E)$
2: **for all** peaks $p(i) \in S$ **do**
3: Create new vertex v with properties $i_v = i \wedge m_v = m_p$
4: $V \leftarrow V \cup v$
5: **end for**
6: **for all** vertices $v_i \in V$ **do**
7: $d \leftarrow 0$
8: $j \leftarrow i$
9: **while** $d \leqq ld \wedge j < |V|$ **do**
10: $j \leftarrow j + 1$,
11: $d \leftarrow |m_{v_j} - m_{v_i}|$
12: **if** $d \leqq ld$ **then**
13: Create new arc $e_{x,y}$ where x and y are indices of tail and head vertex respectively
14: $Weight(e) \leftarrow d$
15: **if** $i_{p_i} < i_{p_j}$ **then**
16: $e.x \leftarrow i$, $e.y \leftarrow j$
17: **else**
18: $e.x \leftarrow j$
19: $e.y \leftarrow i$
20: **end if**
21: $E \leftarrow E \cup e$

end if
end while
end for
$G'(V', E') \leftarrow TopologicalSort(G(V, E))$
for all $v'_i \in V'$ such that $|OutE_{v'_i}| > 0$ **do**
  Find $e_{i,j} \in OutE_{v'_i}$ such that $\nexists e_{k,l} \in OutE_{v'_i}$ such that $w_{k,l} < w_{i,j}$
  if $i_{v'_i} < i_{v'_j}$ **then**
    $i_{v'_i} \leftarrow i_{v'_i} + i_{v'_j}$ {merge intensities}
    $i_{v'_j} \leftarrow 0$
    $OutE_{v'_i} \leftarrow \emptyset$
    $OutE_{v'_j} \leftarrow \emptyset$
  else
    $i_{v'_j} \leftarrow i_{v'_j} + i_{v'_i}$ {merge intensities}
    $i_{v'_i} \leftarrow 0$
    $OutE_{v'_i} \leftarrow \emptyset$
    $OutE_{v'_j} \leftarrow \emptyset$
  end if
end for

The graph creation has almost linear $O(N)$ time complexity (where N is the number of peaks in spectrum) because very few peaks are closer than 0.3 Da (dependent on mass spectrometer resolution) to each other. Prior to peaks merging, a topological sort must be performed. The topological sort algorithm creates a linear ordering of the vertices such that if an edge $e(u, v)$ appears in the graph, then v comes before u in the ordering. The time complexity of topological sort is $O(V + E)$. The next step is to merge sorted vertices beginning with the lowest edge weight. Time complexity of this operation in the worst case is $O(V \cdot E)$. This is then the complexity of

the whole algorithm.

5.4 The Algorithm "Make Equidistant Spectrum"

If we want to consider an MS/MS spectrum as a signal in time domain, it is necessary to convert the spectrum into a signal with equal distances. The algorithm screens through all peaks in spectrum. If the m/z value of two peaks differs by a value less then required step distance, the peak with lower intensity is deleted. All peaks need to have m/z values as:

$$m_1 + f \cdot d \tag{5.1}$$

where m_1 is the m/z value of the first peak in spectrum, d is required step distance between peaks and f is an multiplication factor. For absent m/z values new peaks have to be added with an intensity set to 0.

The time complexity of the algorithm is linear to the number of peaks. The space complexity is $O(\frac{N+R}{D})$ where N is the number of peaks and D is the chosen distance between two signals, and R indicates all imaginary peaks with intensity 0 which had to be added to form a legal signal in time domain. This value is strongly dependent on the spectrum quality. In the spectra with a huge number of noise peaks, this value can be very small but still considerable because real spectra (even if they are very noisy) do not have a peak registered on every 0.3 Da (which is an example of the lowest distance for the peak merging described in the last section).

Make Equidistant Spectrum

Require:

	S	Ordered set of peaks (from merged spectrum)
	e	Equidistant step distance between peaks
	d	Isotope distance between peaks
	t	Mass tolerance

1: $i \leftarrow 0$
2: $size \leftarrow NumberOfPeaks(S)$
3: Create emty set of peaks S'
4: **for all** $0 < i < size$ **do**
5: $j \leftarrow i + 1$
6: $currMass \leftarrow m_i$
7: $nextMass \leftarrow m_j$
8: $intensity \leftarrow i_i$
9: **if** $|nextMas - currMass| < e \wedge i_j > i_i$ **then**
10: $intensity \leftarrow i_j$
11: **end if**
12: $S' \leftarrow S' \cup k$ such that $m_k = currMass \wedge i_k = intensity$
13: **while** $|nextMass - currMass| > t \cdot d$ **do**
14: $currMass \leftarrow currMass + e$
15: $S' \leftarrow S' \cup k$ such that $m_k = currMass \wedge i_k = 0$
16: **end while**
17: **end for**
18: $S' \leftarrow S' \cup k$ such that $m_k = m_{size} \wedge i_k = i_{size}$

5.5 The Algorithm "Calculate Isotope Pattern"

The intensity patterns in isotope clusters become complicated with large fragment masses but still can be exactly calculated [57, 58, 59, 60, 61]. Given the large number of potential peptide fragment sizes and sequence possibilities, the computational time for taking into account the exact isotopic patterns is too high for a background analysis program. As a computational shortcut for calculating the intensities of expected multiply charged peak cluster, the Wehofsky's polynomial approximation [39, 62] was used, where the relative intensity of the n^{th} isotope variant peak (in a pattern of peaks; $N \leqq 7$, $k = 6$ the order of expansion) is:

$$I(n, M) = A(n) + \sum_{j=1}^{k} B_j(n) \cdot M^j \tag{5.2}$$

The intensity patterns have been tabulated with an accuracy of 100 Da (m/z window width). M is the mass corresponding to the first, monoisotopic peak (n=1) in the current m/z window. The relative intensity of this peak is assumed to be 1. $A(n)$ and $B_j(n)$ are fitting parameters taken from Wehofsky's work [39, 63]. Depending on the charge state z, the mass-to-charge-ratio distance between peaks in the pattern is $\frac{1}{z}$ Da and the pattern length is $\frac{N-1}{z}$ Da.

Calculate Isotope Pattern

Require:	N	Number of peaks in the isotope peak cluster
	k	Order of expansion
	M	Set of m/z windows
	A	Set of fitting parameters

B Set of fitting parameters

1: **for all** $m \in M$ **do**
2: $n \leftarrow 1$
3: **while** $n \leqq N$ **do**
4: $Intensity(n, m) \leftarrow A(n)$
5: **for all** j such that $0 < j \leq k$ **do**
6: $Intensity(n, m) \leftarrow Intensity(n, m) + B_j(n) \cdot m^j$
7: **end for**
8: **end while**
9: **end for**

5.6 The Algorithm "Dense Spectrum"

Applying this algorithm on the merged (Algorithm 5.3) and equidistant (Algorithm 5.4) spectrum is required by the algorithm "Deconvolute" described in the next chapter. The "Deconvolute" algorithm calculates a correlation coefficient between experimental and theoretical signals. To achieve high correlation both signals were densed by adding artifical peaks.

The mass window with the length of the target signal (multiply charged isotope peak cluster, Figure 5.6) following each peak is densified with linearly interpolated additional peaks up to the last experimental peak in the window (Figure 5.7). The addition of further peaks (essentially a transformation to a semi-analogue signal) compensates for possible small inaccuracies in resolving the position of isotope-variant peaks by the instrument's software.

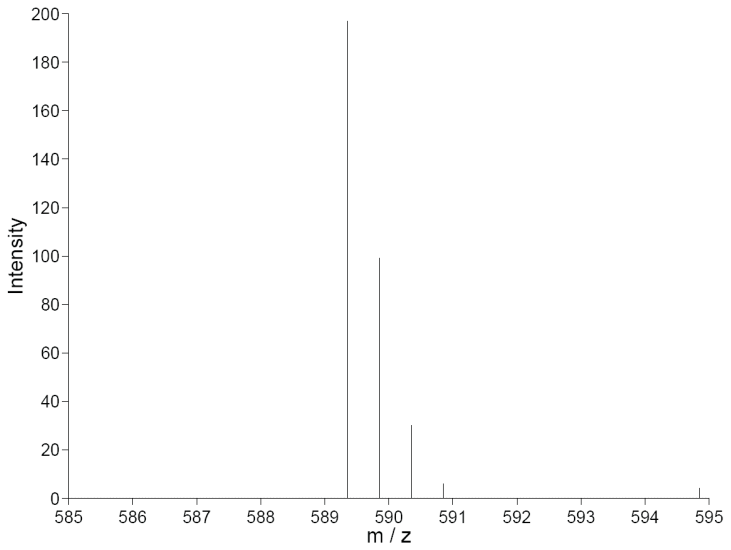


Figure 5.6: Example of a multiply charged isotope peak cluster

Dense Spectrum

Require:	S	Ordered set of peaks
		(from merged, equidistant spectrum)
	d	Distance between isotopic peaks
	t	Mass tolerance
	l	Length of m/z window

1: $i \leftarrow 1$
2: **while** $i \leqq SizeOf(S) - l$ **do**
3: Create empty ordered set of peaks S'
4: $S' \leftarrow$ all peaks $p_n \in S$ such that $|m_{p_n} - m_{p_i}| < l \pm t$
5: **for all** $p(k) \in S'$ such that $i_{p(k)} > 0 \wedge \exists f$ such that $d \cdot f = |m_{p_k} - m_{p_i}| \pm t$ **do**
6: **for all** $p(j) \in S'$ such that $i < j < k$ **do**
7: $i_{p_j} \leftarrow i_{p_i} - \frac{i_{p_i} - i_{p_k}}{k-i}$

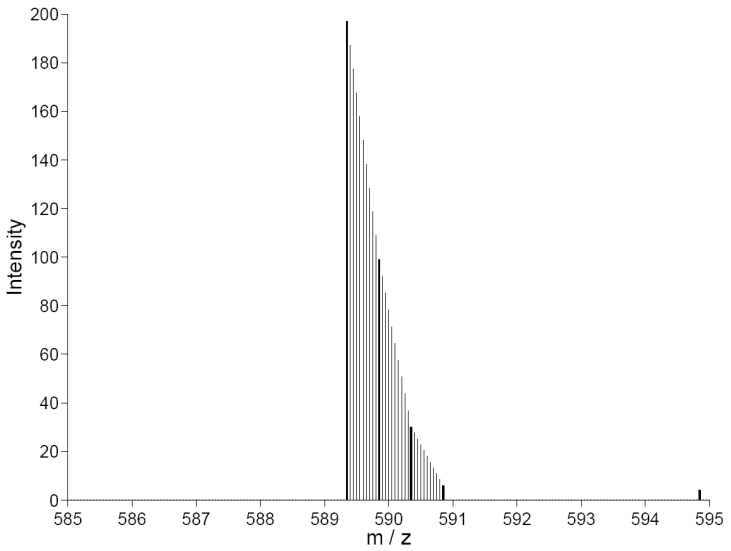


Figure 5.7: Densed multiply charged isotope peak cluster

8: **end for**

9: **end for**

10: $i \leftarrow k$

11: **end while**

The algorithm screens for peaks within the isotope cluster length l. If the distance between found peaks in one cluster of peaks is approximately identically to the required distance between isotopic peaks d, new peaks with interpolated intensity between original peaks are added.

The time complexity of the densification algorithm is $O(N \cdot \frac{I}{d})$ where N is the number of signals in equidistant spectrum, I is the number of found peak clusters that could form a multiply charged isotope peak cluster and d is the distance between isotope peaks for the considered charge state.

5.7 The Algorithm "Deconvolute Spectrum"

Although ionization techniques such as ESI have the advantage of shifting heavy ions into lower, detectable mass-over-charge ranges by generating multiply charged fragment ions [34], they can pollute the spectrum by causing replicates of otherwise identical peptide fragments with different charge states. It is expected that the multiply charged signals do not occur as monoisotopic peaks but as isotope peak clusters. The distance between the peaks in a peak cluster depends from the charge state of the considered fragment ion. For the purpose of spectrum interpretation, peak replicates originating from different charge states have to be unified. This includes transforming the monoisotopic peaks with higher charge states into singly charged monoisotopic peaks as well as removing heavy isotope peaks with higher charge state. Removal of singly charged heavy isotope peaks is described in section 5.9.

The relative spectral intensities of isotope-variant peaks in a cluster are determined by the natural isotope distributions of carbon, hydrogen, oxygen, nitrogen and sulphur, the predominant chemical elements in peptide fragments. This a priori known form of the intensity pattern from multiply charged replicates was used in this work to identify multiply charged peaks in the measured spectrum by correlational analysis.

Although the method for removal of latent periodical noise (including singly charged isotope clusters) described in section 5.9 detects in some cases also the multiply charged isotope peaks, the best results are obtained by applying a dedicated procedure for finding doubly- and triply-charged isotope peak clusters. A new algorithm is developed for this purpose. The algorithm is quite robust with respect to inaccuracies in the experimental resolution of isotope clusters due to two artifices in processing the mass spectrum:

(i) the merging of small peaks very close to major intensities,

(ii) the procedure of interpolated peak densification in the mass range of comparison with the predefined pattern.

The algorithm includes several steps (Figure 5.8). Prior to spectrum analysis, the general form (the etalon) of isotope cluster patterns is pre-computed for double- and triple-charged fragments (Equation 5.1). Higher charge states can be easily added by pre-calculating the isotope pattern for the considered charge state. It is also possible to process negative charge states. This feature demands a simple correction in the calculation of the precursor mass and m/z values if necessary.

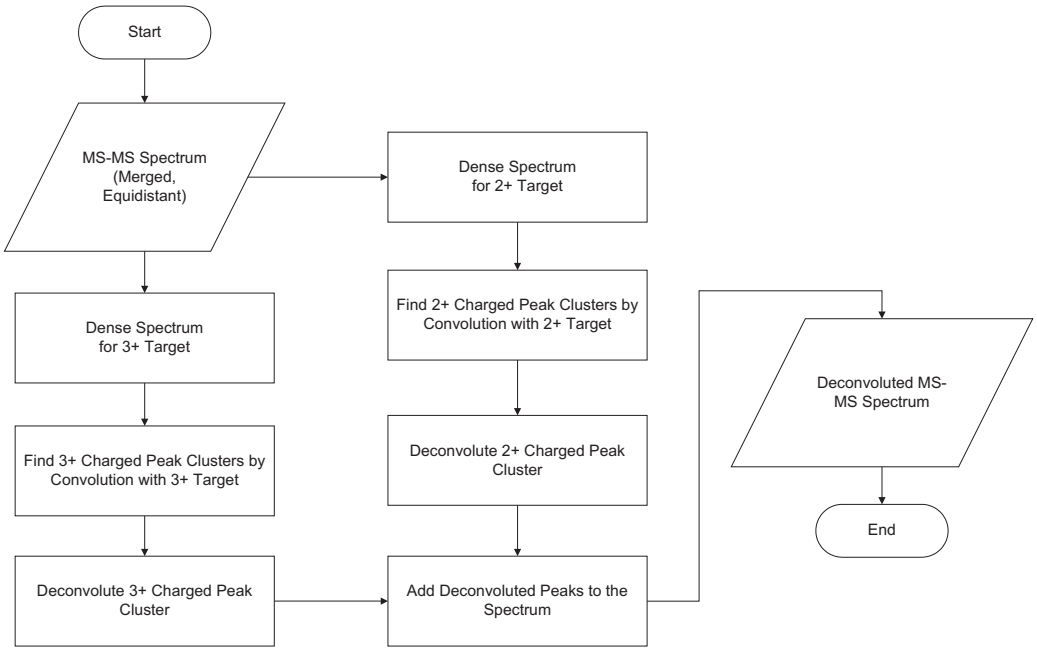


Figure 5.8: Determination of multiply charged replicates with correlation analysis.

To obtain high correlation coefficients the pattern of the etalon has to be densed for a particular charge state with $\frac{1}{d} \cdot \frac{N-1}{z} - N + 1$ additional peaks

in total (with the mass/charge-ratio step, d, as defined in section 5.4) where their intensity is linearly interpolated from the two surrounding pattern-defining peaks with masses $M + \frac{n-1}{z}$ and $M + \frac{n}{z}$ (Figure 5.6).

The spectrum is divided into windows of peaks (for example 100 Da wide) in which peak groups are searched that could form an isotope cluster. Dividing the spectrum in windows is important because of the pattern dependence on m/z vaules. The isotope pattern search criteria are the distance between peaks and the shape of peak groups. The exactly calculated shape of groups of peaks is used as a target signal for the convolution in a particular window. If a group of peaks roughly shows characteristics of an isotope peak cluster (for example the distance between peaks is ≈0.3Da and first peak is larger than the second one in the 300-399 Da m/z window) densification is performed for this group of peaks. The intensities are interpolated and a new peaks are added on every (chapter 5.6).

The next step is the calculation of the correlation coefficient between the spectrum and target signal for considered window and charge state. Every peak is considered as possible start point of an isotope cluster. The correlation coefficient is calculated as follows:

$$r = \frac{N \cdot \sum_{i=1}^{N} X_i \cdot Y_i - \sum_{i=1}^{N} X_i \cdot \sum_{i=1}^{N} Y_i}{\sqrt{[N \cdot \sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2][N \cdot \sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2]}} \tag{5.3}$$

where N is number of signals in the target, X_i and Y_i are the intensities of the experimental peaks and the intensity of theoretically calculated peaks respectively. The time complexity of the correlation algorithm is then $O(N \cdot M)$ with N being the number of peaks in the spectrum and M number of signals in the target signal (isotope peak cluster for considered charge state

and window). The correlation coefficient is calculated for both 2^+ and 3^+ charge states. If the coefficient is higher than a user-defined threshold (for example 0.95) the group of peaks is considered as a multiply charged isotope peak cluster. If both coefficient for 2^+ and 3^+ charge are higher than the threshold, the charge state is determined according to the higher correlation coefficient.

If a multiply charged isotope cluster is found in the spectrum and confirmed with a high correlation coefficient, the original peaks that belong to the isotope cluster are removed from spectrum, and instead of them, a single peak with charge state 1^+ is calculated from the original peaks and added to the spectrum. The m/z value of the first peak in a multiply charged isotope cluster can be represented as $K_i = \frac{M+i \cdot m}{i}$, where i is the charge state, M is the mass of the peptide without charge carrier and m is the mass of the charge carrier (in most cases the proton). The m/z value of the singly charged monoisotopic peak is calculated as $M + m = i \cdot K_i - (i-1) \cdot m$.

An example of a recognized isotope peak cluster is shown in Figure 5.9. The signals above the x-axis are peaks from an MS/MS spectrum. Below the x-axis are corresponding correlation coefficients. In this example, only the first peak is recognized as a monoisotopic peak (manifested by correlation coefficient higher as the defined threshold).

5.8 The Algorithm "Median filter"

For power spectra smoothing, required by the algorithm "Deisotope spectrum" (section 5.9) a special median filter has been developed. Typically, the power spectrum of a good MS/MS spectrum is quasi-periodic. The length of this period is determined with another Fourier-transformation, where the

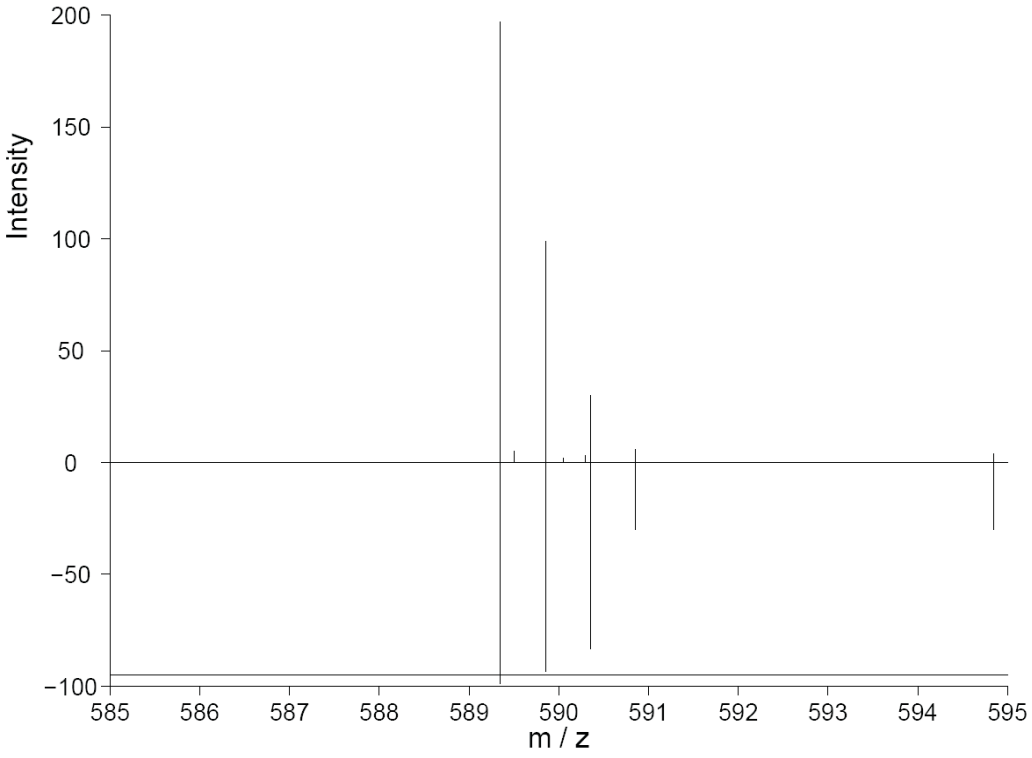


Figure 5.9: Determination of multiply charged replicates with correlation analysis

power spectrum was considered as a signal in the time domain. Determination of the periodicity fails if the power spectrum shows multiple of the base frequency (see section 5.9). This can be bypassed by smoothing the power spectrum before applying the second Fourier-transformation. For this purpose the median filtering is one of standard algorithms [56, **?**], replacing all vaules in the signal by the median from certain range. A median is a number dividing the higher half of a sample from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one.

The median filter is designed by creating two self-balancing binary search trees (AVL trees), the left and the right AVL tree. AVL trees are very

effective data structures for retrieving smallest or highest elements. From N signals, those which are smaller than the median are stored in the left tree and the large ones in the right. The median was held as a value higher than the highest element of the left AVL tree and less than the minimum of the right AVL tree. In this way the time complexity of the median filter algorithm is O(Nlog$N$).

Median filter

Require:	*S(N)*	Vector of signals to be smoothed by median filter
	N	Size of the vector of signals
	F	Median filter size

1: Create emty set MFSthis is the median filtered output signal
2: $cnt \leftarrow 0$
3: $median \leftarrow S(0)$
4: Create empty AVLTree L
5: Create empty AVLTree R
6: **for all** signals $s_i \in S$ such that $0 < i < \frac{F}{2} \wedge i < N$ **do**
7: **if** $s_i < median$ **then**
8: $L \leftarrow L \cup s_i$
9: **else**
10: $R \leftarrow R \cup s_i$
11: **end if**
12: **end for**
13: $LastRightBound \leftarrow N - 1$this is the index of the last signal in S
14: $RightBound \leftarrow LastRightBound$this is the index of the last signal in the neighborhood, the neighborhood width is always $\frac{F}{2}$
15: **while** $RightBound < N + 1$ **do**

16: **for all** $s_i \in S$ such that $LastRightBound \geqq i < RightBound$ **do**
17: **if** $s_i < median$ **then**
18: $L \leftarrow L \cup s_i$
19: **else**
20: $R \leftarrow R \cup s_i$
21: **end if**
22: **end for**
23: **if** $cnt > \frac{F}{2}$ **then**
24: $j \leftarrow cnt - \frac{F}{2} - 1$
25: **if** $s_j < median$ **then**
26: $L \leftarrow L \setminus s_j$
27: **else**
28: $R \leftarrow R \setminus s_j$
29: **end if**
30: **end if**$median = \text{BalanceTrees}(median, L, R)$
31: $MFS \leftarrow MFS \cup median$
32: $cnt \leftarrow cnt + 1$
33: $LastRightBound \leftarrow RightBound$
34: $RightBound \leftarrow \min(RightBound + 2, cnt + \frac{F}{2} + 1$
35: **end while**
36: **for all** $s_i \in S$ such that $(n - 1 - \frac{F}{2}) \leqq i < n$ **do**
37: **if** $s_i < median$ **then**
38: $L \leftarrow L \cup s_i$
39: **else**
40: $R \leftarrow R \cup s_i$
41: **end if**
42: **end for**

43: **while** $cnt < n$ **do**
44: **if** $s_{cnt} < median$ **then**
45: $L \leftarrow L \cup s_{cnt}$
46: **else**
47: $R \leftarrow R \cup s_{cnt}$
48: **end if**
49: **if** $cnt > \frac{F}{2}$ **then**
50: **if** $s_{cnt} < median$ **then**
51: $L \leftarrow L \setminus s_{cnt}$
52: **else**
53: $R \leftarrow R \setminus s_{cnt}$
54: **end if**
55: **end if**$median = \text{BalanceTrees}(median, L, R)$
56: $MFS \leftarrow median$
57: $cnt \leftarrow cnt + 1$
58: $LastRightBound \leftarrow RightBound$
59: $RightBound \leftarrow \min(RightBound + 2, cnt + \frac{F}{2}) + 1$
60: **end while**
61: return MFS

BalanceTrees

Require:	$median$	current median value
	L	Left tree
	R	Right tree

1: **while** $|L| > |R| + 1$ **do**
2: $R \leftarrow R \cup median$

3: $median \leftarrow HighestElementOf(L)$
4: $L \leftarrow L \setminus median$
5: **end while**
6: **while** $|R| > |L| + 1$ **do**
7: $L \leftarrow L \cup median$
8: $median \leftarrow SmallestElementOf(R)$
9: $R \leftarrow R \setminus median$
10: **end while**
11: return $median$

5.9 The Algorithm "Deisotope spectrum"

Due to the natural isotopic distribution, masses in mass spectrometer are not detected as single peaks (monoisotopic peaks), but as groups of peaks (peak cluster) with different intensity and defined mass difference. A mass spectrum of any organic compound will usually contain a small peak of one mass unit (Da) greater than the apparent molecular ion peak (M). This is known as the M+1 peak and originates due to the presence of carbon-13 atoms (^{13}C-isotope). Natural occurrence of the ^{13}C-isotope is ≈1.1%. A molecule containing one carbon atom will be expected to have an M+1 peak of approximately 1.1% of the intensity of the M peak as 1.1% of the carbon atoms will be carbon-13 rather than carbon-12. If an isotope cluster is singly charged, the distance between the peaks is 1Da.

It should be expected that isotope clusters are the source of latent periodicity in the signal that should be visible in form of maxima in the frequency spectrum of the signal. A test was performed to prove this assumption. In Figure 5.10, an original peptide MS/MS spectrum is depicted.

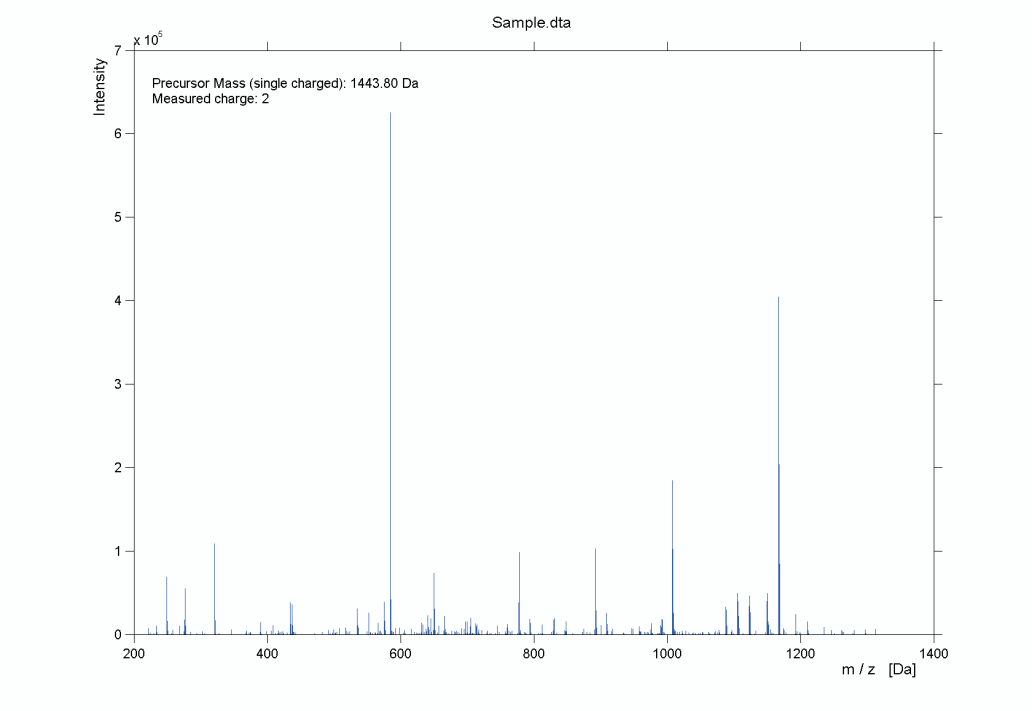


Figure 5.10: Example of an MS/MS spectrum

Its corresponding power spectrum (PS)[56] is shown in Figure 5.11.

The power spectrum is calculated as:

$$PS = \frac{Z - \bar{Z}}{n} \tag{5.4}$$

where Z is Fourier-transformed MS/MS spectrum, $\bar{Z}$ is its complex conjugate and n is the number of signals.

If we extract only the peaks that are interpretable by a database search program (for example Mascot [30]) we get an artificial spectrum such as it is shown in Figure 5.12. The spectrum has no repeatable signals and this fact is confirmed in its power spectrum 5.13. The same spectrum with artificially added isotope peaks is shown in Figure 5.14. This artificial MS/MS spectrum exhibits latent periodicity in its Fourier transforms 5.15. Thus, disappear-

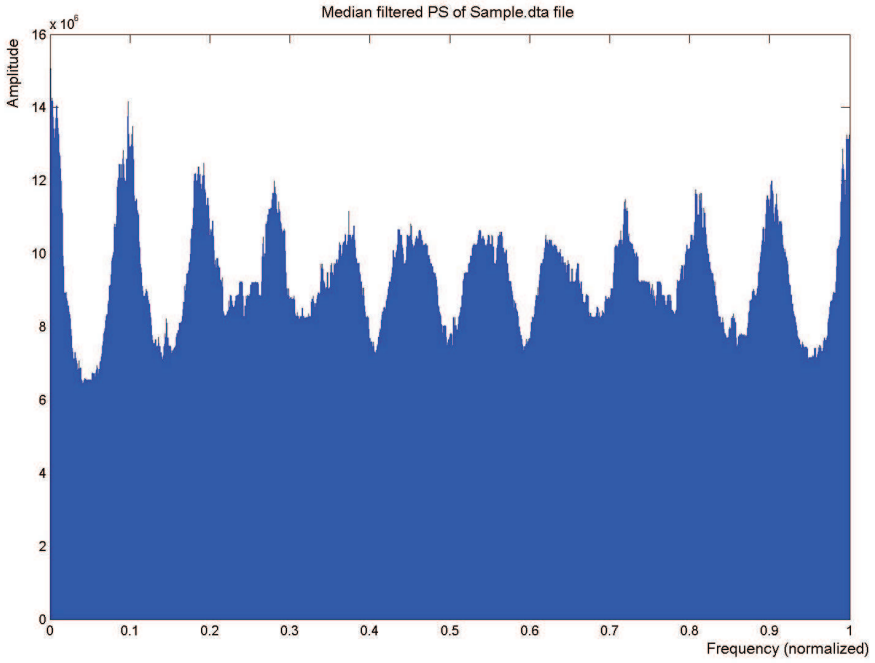


Figure 5.11: The same MS/MS spectrum in the frequency domain

ance of isotope clusters correlates with dampening of the prominent periodic spectral component in the Fourier transform.

Because the singly charged isotope peak clusters have the repetition of signals as in the case of multiply charged isotope peak clusters, theoretically, the method described in chapter 5.7 could be used to detect the singly charged clusters as well. Correlation of the measured MS/MS spectrum with pre-calculated isotopic intensity distributions is efficient only for multiply charged peak cluster detection. Singly charged peak clusters cannot be reliably detected with the method described in the previous chapter since the probability of finding additional, unrelated peaks in the spectrum with a distance of 1 Da is high. Therefore, correlation analysis with pre-defined patterns is not really useful for deisotoping. But if we treat an MS/MS

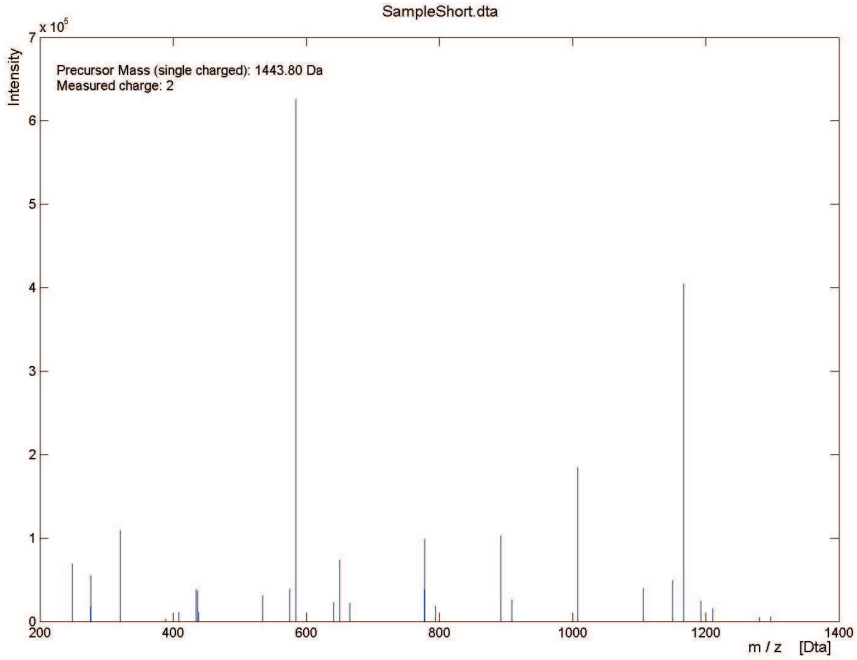


Figure 5.12: MS/MS spectrum containing only interpretable peaks

spectrum as a set of signals in the time domain where the mass-over-charge axis is the analogue of time and the intensity of each peak in the MS/MS spectrum as the intensity of a signal at a certain time, we can consider the single-charged peak signals as periodical function (with periodicity of ≈1 Da for singly charged peaks). This periodical function in the time domain results in a power spectrum in the frequency domain where the reoccurring elements can be much easier recognized.

Besides isotope variants, there can be other sources of spectral contamination with latent periodicity, for example from the electronic detection system or from accompanying chemical polymer contaminants such as silanes, etc. Re-occurring signals at quasi-constant mass shifts can be seen in the frequency domain, i.e. as characteristic reoccurrences of high amplitudes at

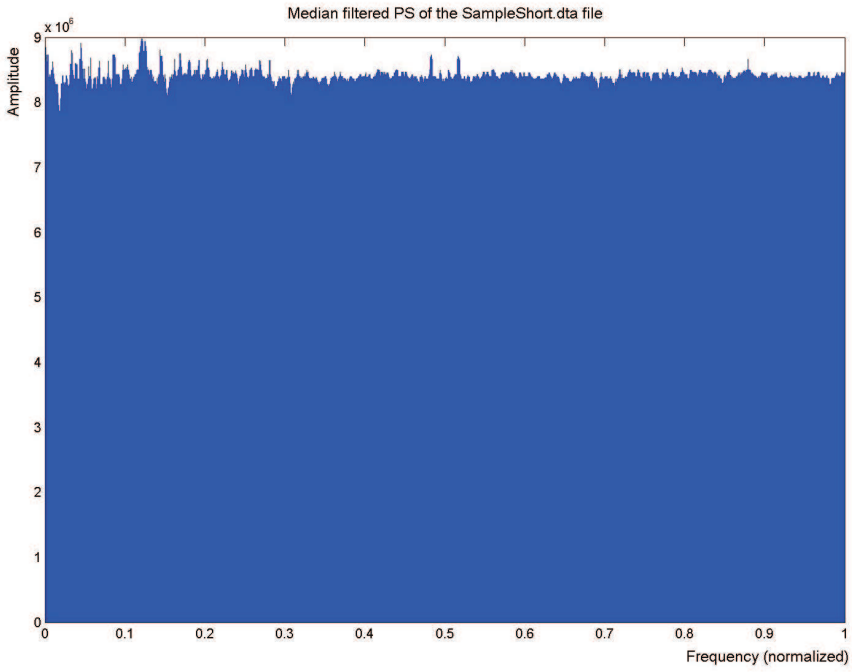


Figure 5.13: Spectrum containing only interpretable peaks looked in frequency domain

multiples of a base frequency f_B in the Fourier transform of the tandem mass spectrum. A similar periodicity analysis has been previously proposed for the detection of chemical background in MS fingerprints of small organic or inorganic compounds [64].

Converting to the frequency domain, the discrete Fourier transform Y of the MS/MS spectrum S is found by taking the N-point fast Fourier transform $Y = FFT(S, N)$. The value N is calculated as N^{n+1} , where n is $\lceil \log_2(\frac{x_{max}-x_{min}}{0.05}) \rceil$. The values x_{max} and x_{min} are the largest and the smallest mass-over-charge values in the spectrum respectively. The power spectrum, a measurement of the power at various frequencies, is calculated according equation 5.2. Typically, the power spectrum of a good MS/MS spectrum is quasi-periodic 5.16.

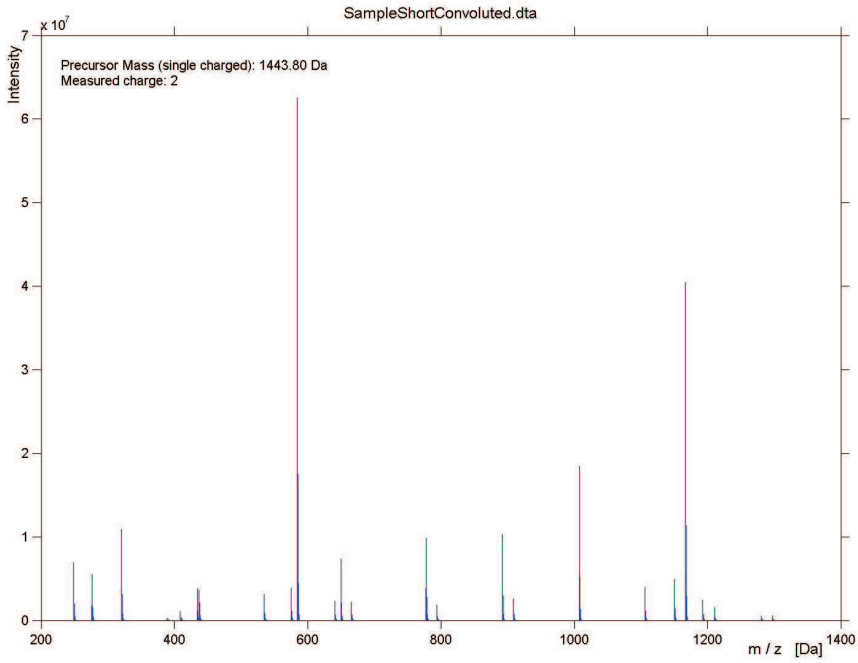


Figure 5.14: MS/MS spectrum containing interpretable peaks with artificially added heavy isotope peaks

The length of this period (the base frequency fB) is determined with another Fourier-transformation, where the power spectrum was considered as a signal in the time domain (Figure 5.17, called PSPS-graph below).

In order to remove the reoccurring elements from the power spectrum, a multi-band reject filter has to be introduced for each MS/MS spectrum. There exist many "standard" modeling techniques using a digital filtering approach based on different spectral estimation methods [74]. Filter design functions such as $yulewalk$, $invfreqz$, and $cremez$, are available. The selection of a method depends on the available response data and target criteria of the design. A multi-band reject filter is created by the Yulewalk method of autoregressive moving average (ARMA) spectral estimation [65]. Yulewalk designs recursive infinite impulse response (IIR) digital filters using a least

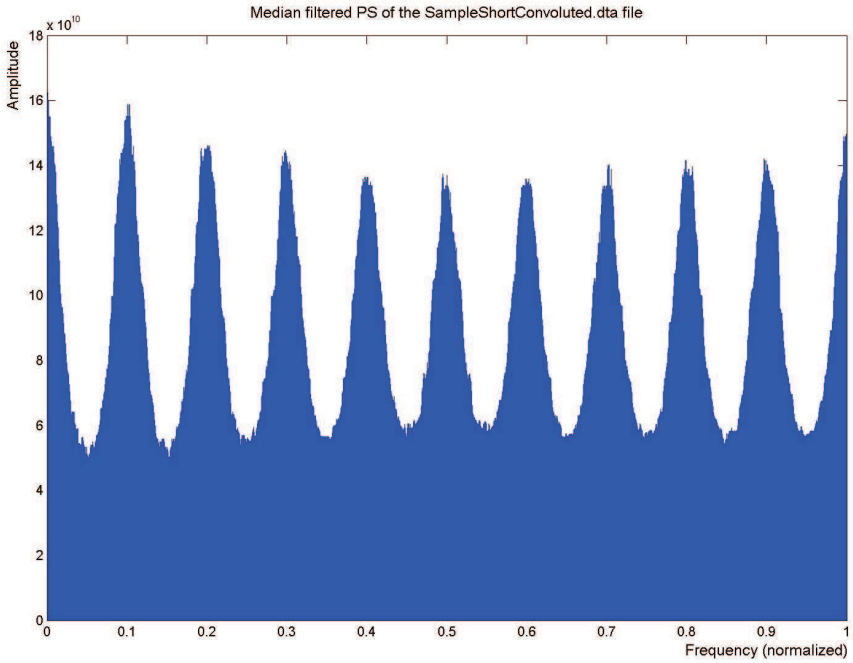


Figure 5.15: Interpretable peaks and artificial added heavy isotope peaks looked in frequency domain

squares fit to a specified frequency response:

$$[b, a] = yulewalk(n, f, m) \tag{5.5}$$

The Yulewalk algorithm returns row vectors b and a containing the $n+1$ coefficients of the order n IIR filter whose frequency-magnitude characteristics approximately match those given in vectors f and m: f is a vector of frequency points, specified in the range between 0 and 1, where 1 corresponds to half the sample frequency (the Nyquist frequency). The first point of f must be 0 and the last point 1, with all intermediate points in increasing order. Duplicate frequency points are allowed, corresponding to steps in the frequency response. m is a vector containing the desired magnitude response

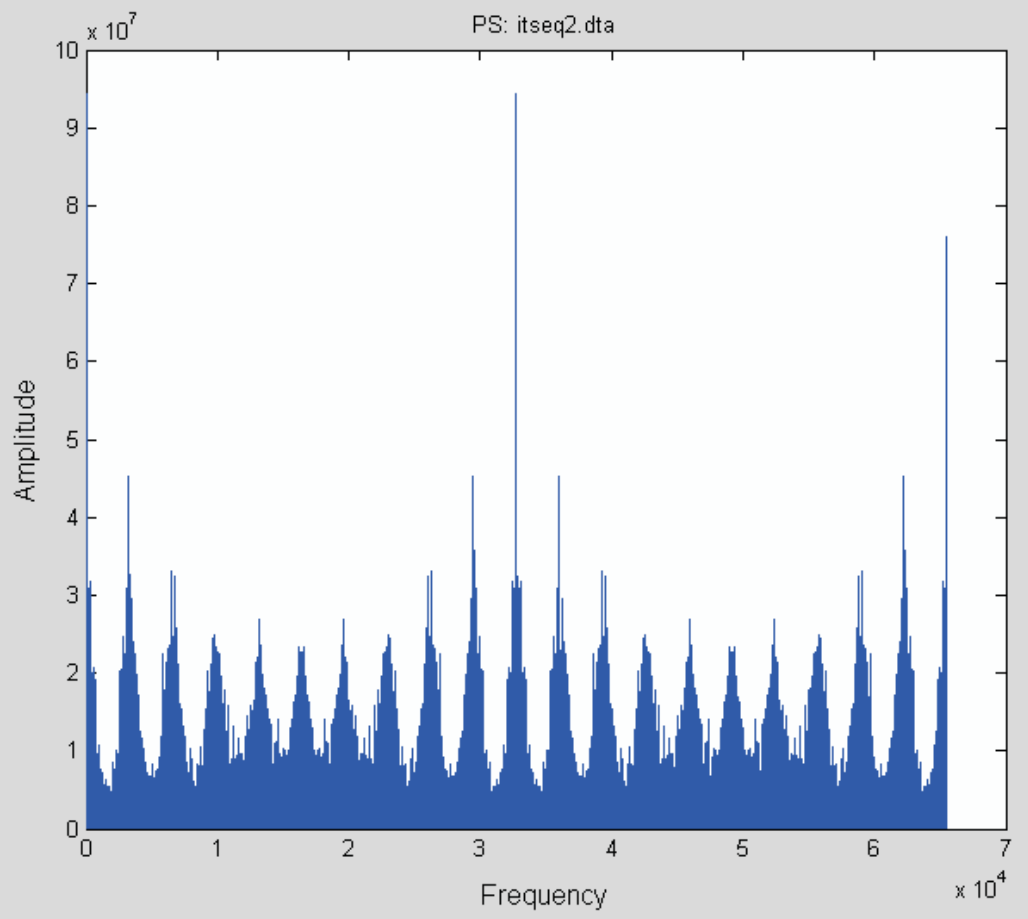


Figure 5.16: Power spectrum of an MS/MS spectrum showing periodical amplitudes

at the points specified in f. f and m must be the same length. The Yulewalk algorithm's time complexity is not bound to the number of signals in MS/MS spectra but on the size of frequency vector which is irrelevant compared to the size of the spectrum.

Frequencies required by the Yulewalk algorithm are calculated by applying a median filter to the power spectrum (over 300-500 discrete data points, see section 5.8) and by computing a second power spectrum (PSPS-graph) in order to obtain the most prominent frequency of the first power spec-

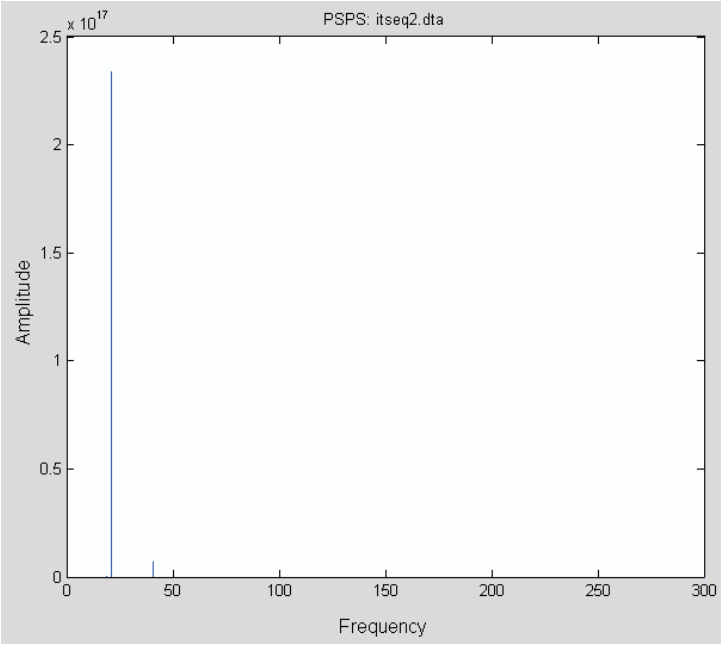


Figure 5.17: PSPS spectrum showing the periodicity of the first power spectrum

trum (PS-graph). The method of taking the most prominent frequency from the PSPS-graph is called *rigorous periodicity detection* (including some bad spectra detection). In some cases described in section 5.11.1, this periodicity detection is not possible (absence of a clear maximum in the PSPS spectrum). In such cases, a *soft detection method* can be used. Soft periodicity (and bad spectra) detection consists of calculating the coefficient of dispersion (see the section 5.11.1) for every frequency in the PSPS-graph and of selecting the optimum.

With the calculated frequency of the power spectrum, the Yulewalk can be performed. The result of the Yulewalk algorithm is a recursive IIR digital filter [65, 67] described by the numerator and denominator coefficient vectors. For each MS/MS spectrum, a new filter is created and a spectrum is filtered in the time domain [67]. The time complexity of this operation is $O(Z_i \cdot N)$ where N is the number of signals in the equidistant MS/MS spectrum and

Z_i is a vector of length $\max(length(a), length(b)) - 1$. The coefficients a and b are the numerator and denominator coefficients of the IIR filter (see references [65, 67]). The length of Z_i is much smaller than the size of an MS/MS spectrum and it depends on the detected periodicity in the PS graph.

Applying the multi-band reject filter on an MS/MS spectrum reduces the intensity of all signals in time domain. The most affected peaks are latent periodic noise peaks (including isotope peaks). The peaks that have lost on their intensity more than a user-defined value (for example 99.9%) are marked for removal from the original spectrum (Figure 5.18).

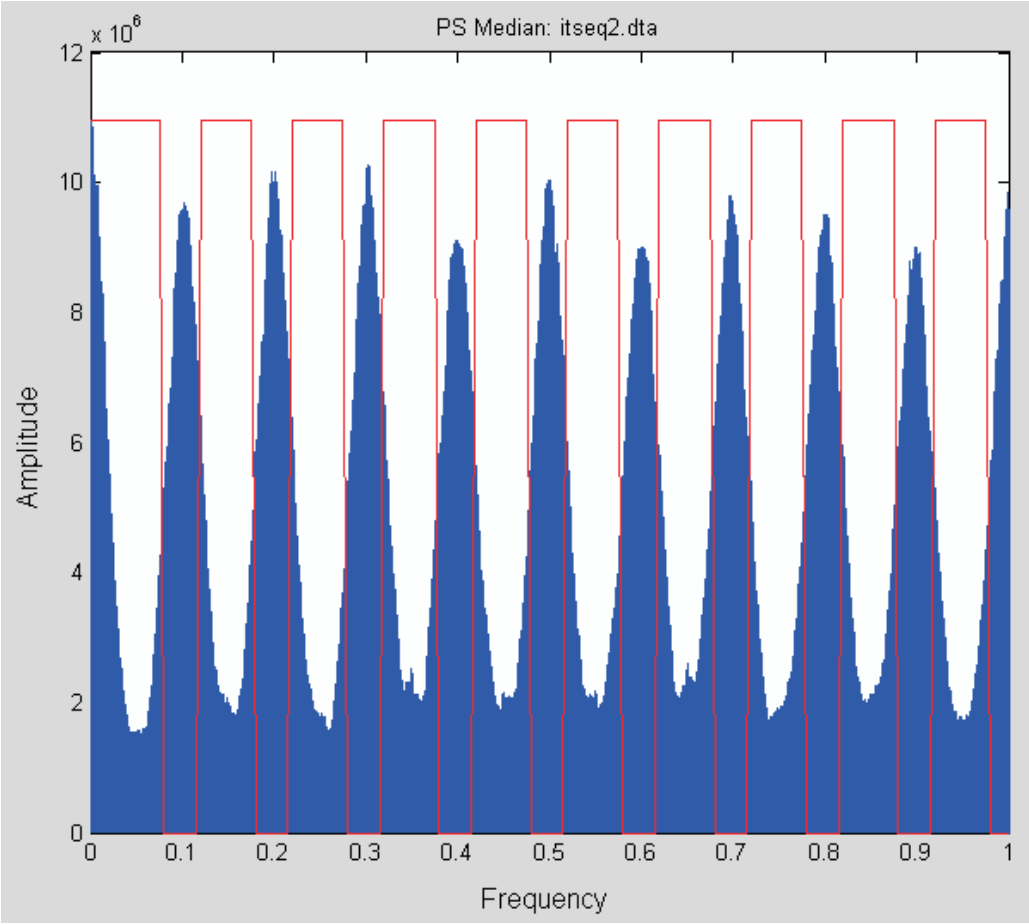


Figure 5.18: Multiband-reject filter overlayed on the PS

After filtering, the recovered MS/MS spectrum might contain some signals with negative intensity or some new signals with positive intensity. These two types of signals are corrected to zero. Additionally, some signals occurring with positive intensities both in the original raw spectrum and the recovered spectrum have lost considerable intensity in the later (threshold of 95%; this number should be higher for very clean and regular spectra). The result of applying the multi-band reject filter on the raw spectrum is shown in Figure 5.19. The intensity decrement is different for each peak. Only peaks that were periodical replicates have lost the most intensity (depicted in green in Figure 5.19).

The peaks which have lost on intensity more than an empirically determined value (for example more than 99.99%) are marked for removal from original MS/MS spectrum. Examination of exemplary spectra has shown that suppression of latent periodicities in the MS/MS spectrum effectively also removes peaks originating from heavy isotope peaks in isotope peak clusters (Figure 5.19).

It should be noted that this algorithm is developed only for marking peaks for deletion. A spectrum which has lost some frequencies in the PS can not be used for further analysis, because by applying the multi-band reject filter also the monoisotopic peaks have lost on the intensity as well. Transforming the spectrum from frequency domain into time domain and comparison of the decreased intensity mark the heavy isotope peaks. Marked peaks are then deleted from the original spectrum (unmodified in the frequency domain), and this spectrum is then used for further processing and as a final output. This is specially emphasized because a reader could get the impression that the spectrum modified in PS is used as final output. This would be problematic since modifying PS causes modification on all peaks in the time

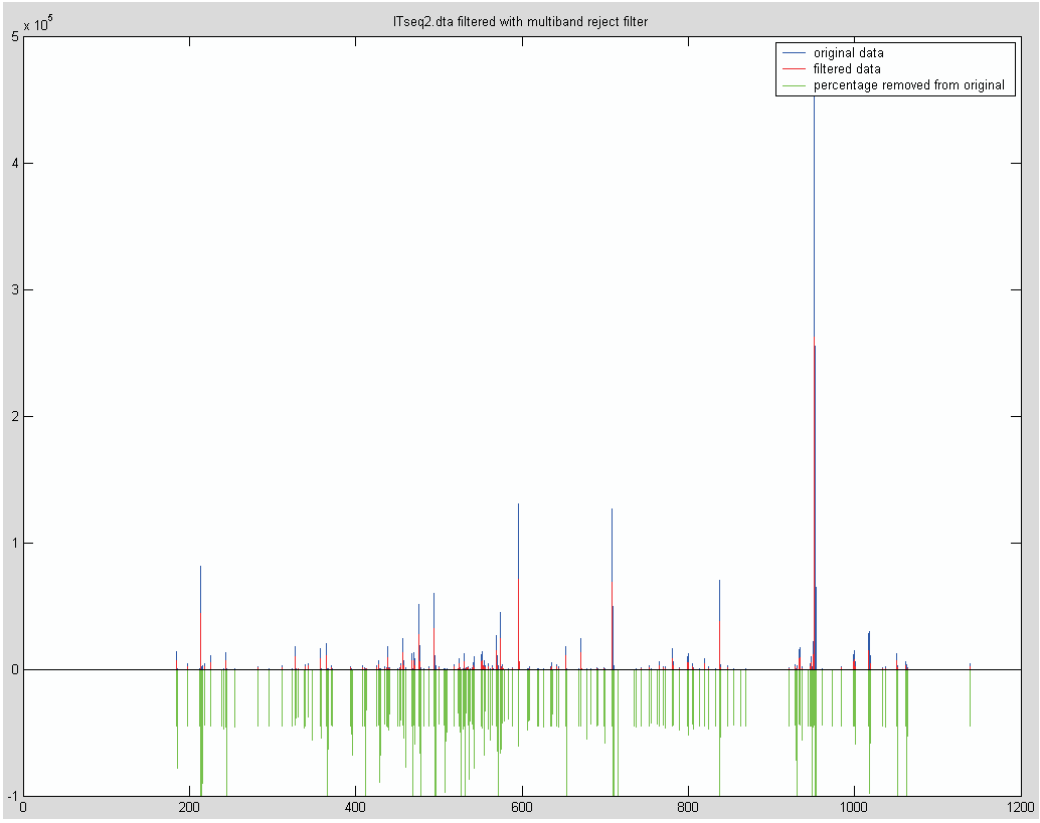


Figure 5.19: An MS/MS spectrum before (blue) and after (red) applying the multiband-reject filter. The percentage of decreased intensity for each peak is shown in green

domain.

In some cases, PS-graphs of dta-files display several, overlaying modes of periodicities. The respective PSPS-graphs have several maxima with similar intensities. If the numerically largest maximum is at very low base frequencies fB (e.g., there is only a few maxima in the PS-graph), the application of the periodical multi-band filter with this fB can lead to severe damage of the MS/MS spectrum. To avoid this problem, intensities in the PSPS-graph are set to zero for low frequencies.

If it is not obvious which frequency to choose from PSPS, i.e. if there are several frequencies visible with almost the same intensity, that spectrum is

probably a bad spectrum and should not be sent to interpretation software. Sometimes this rule doesn't apply and the spectrum is still a good spectrum. This problem is discussed in the section 5.11.1.

A simplified flow chart of the complete deisotoping procedure with *rigorous periodicity detection* is shown in Figure 5.20.

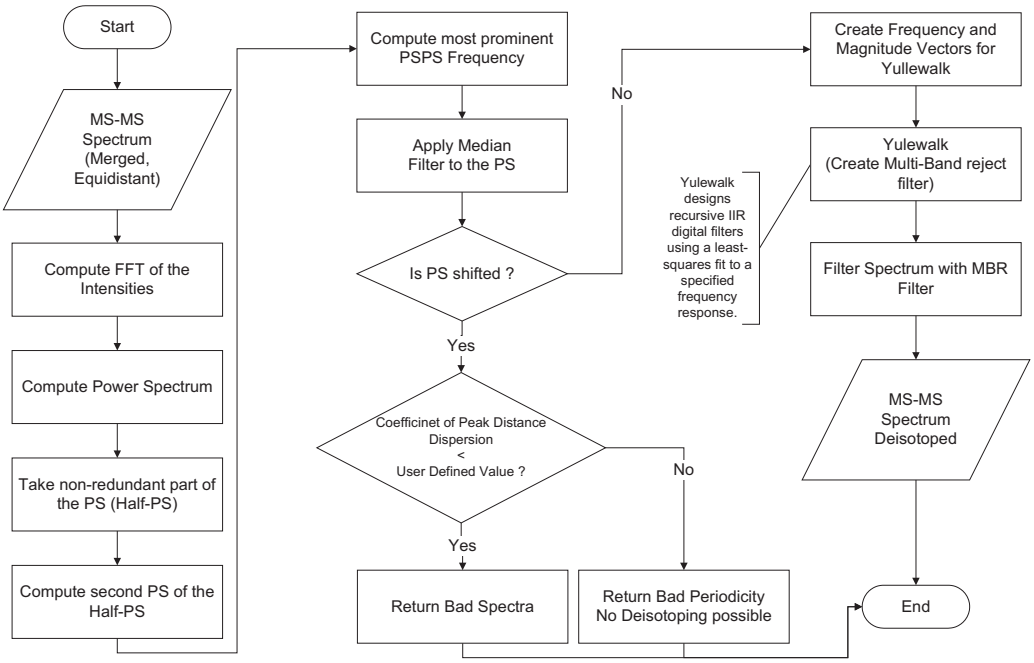


Figure 5.20: Simplified schema of the algorithm "Deisotope spectrum"

The algorithm for *soft* periodicity detection defers from this one in the way the periodicity is obtained from the PSPS spectrum. In the rigorous method, the most prominent frequency is taken from the PSPS graph. In the soft method, all frequencies are checked and the one is taken which has the smallest coefficient of dispersion.

Deisotoping relies on spectra analysis in frequency domain. The signal is transformed from the time domain into frequency domain by applying a Fast Fourier Transform algorithm which has time complexity O(Nlog$N$) [72, 73].

The next step is to calculate the power spectrum (PS)[56] from the signal in frequency domain (in linear time). To investigate an existence of the PS shift (low frequencies with low amplitudes), the PS graph was first smoothed by applying a median filter (section 5.8). The smoothed PS was checked for the frequency shift by calculating the coefficient of dispersion in linear time. If the PS was shifted and the coefficient of dispersion was less than a user defined value (for example ≈3Da) the spectrum was considered as a bad spectrum. If the PS was shifted but the coefficient of dispersion was higher than a user defined value that was an indicator that the periodicity could not be calculated and no decision could be made about the quality of the spectrum.

In order to estimate the periodical frequency amplitudes in the PS, a second PS is calculated (PSPS) considering the PS signal as time domain signal and transforming it into frequency domain. In the rigorous method the periodicity of the PS was determined by the frequency with the highest amplitude from the PSPS. In the case of the soft periodicity detection, a dispersion coefficient has been calculated for all frequencies from the PSPS. This extra layer of complexity did not significantly change the entire time complexity of the algorithm because the number of detected frequency in the PSPS was much lower that the number of points in the power spectrum (highest values observed were 30-40).

The detected frequency periodicity in the PS graph was handed over to the Yulewalk algorithm in the form of frequency and magnitude vectors. The Yulewalk algorithm time complexity is not bound to the number of signals in MS/MS spectra but on the size of frequency vector which is irrelevant compared to the size of the spectrum. The result of the Yulewalk algorithm is a recursive IIR digital filter [65, 67] described by the numerator and denomi-

nator coefficient vectors. For each MS/MS spectrum, a new filter is created and a spectrum is filtered in the time domain [67].

Detection and removal of latent periodical noise including isotope peaks reduces the final number of peaks in an MS/MS spectrum and prevents false interpretation of MS/MS spectra by an interpretation software (see results chapter 7).

5.10 The Algorithm "Remove Random Noise"

Noisy MS/MS spectra suffer from many superfluous peaks densely distributed over the whole mass-over-charge ratio range. These peaks are often the result of electronic noise produced by the mass spectrometer detection system. Assuming that the random noise in an MS/MS spectrum exists as signals of high frequency of occurrence, an IIR low-pass filter [66] can be applied to the spectrum in time domain. A low-pass filter is a filter that passes low frequencies well, but attenuates (or reduces) frequencies higher than the cut-off frequency (Figure 5.21).

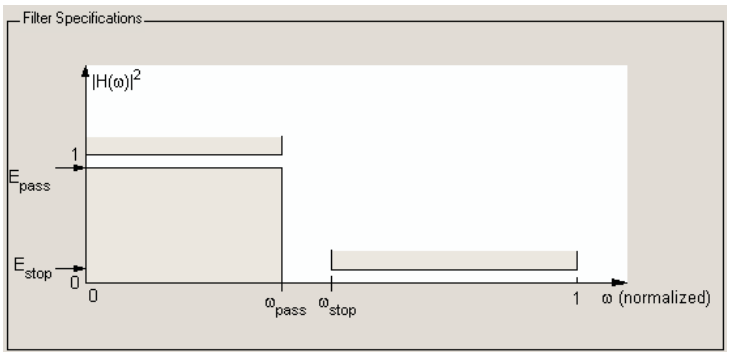


Figure 5.21: Low-pass filter specification

Because of the lack of information about random noise, in order to develop an algorithm for random noise removal, several tests have been carried out

with normalized stop frequency of the filter in the range from 0.5 to 0.9 (Figure 5.22). The best results were obtained with stop frequency 0.8 (see the results section).

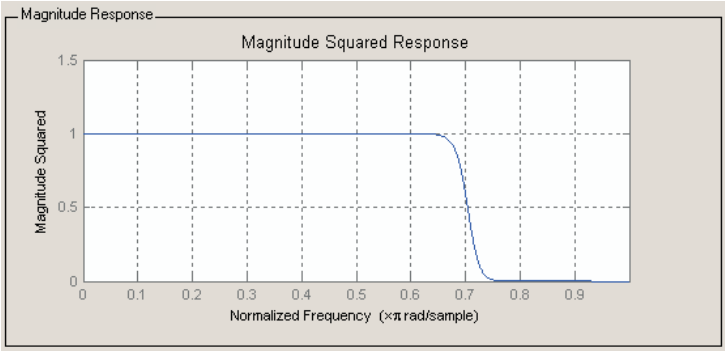


Figure 5.22: Low-pass filter with stop frequency 0.8

As in the case of multi-band reject filter, the application of the low-pass filter reduces intensity of all peaks in a certain amount. An empirical threshold of 99.99% of intensity decrease is applied to mark random noise peaks. The signals which have lost intensity above this threshold are removed from the raw spectrum (Figure 5.23).

It should be noted that the low-pass filtered spectrum was not used in further analysis but only to detect and mark random noise peaks.

5.11 Bad Spectra Recognition

The MS/MS spectra consist mostly of background noise; typically, about 10% of the peaks in a spectrum contribute to the peptide identification and only ≈1% of the spectra are recorded with signals from target protein fragments. Thus, computer resources are mostly spent on analyzing irrelevant data if the identification of the protein with significance is possible within the background at all, a strategy that clashes with limited compute server

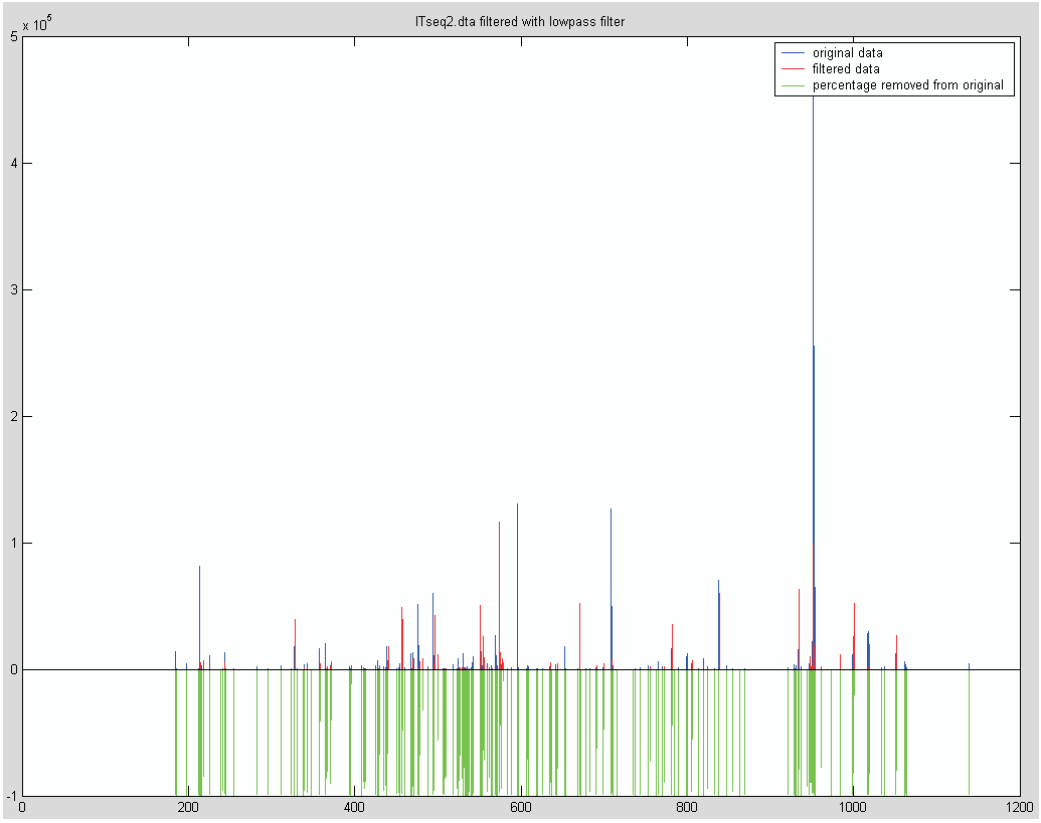


Figure 5.23: An MS/MS spectrum before (blue) and after (red) applying the low-pass filter. The percentage of decreased intensity for each peak is shown in green

capacity in proteomics studies.

With the broad availability of accurate MS/MS instruments with resolution in the order of tenths of a Dalton, automatic background removal procedures before interpretation software application became possible. In this work, several bad spectra recognition strategies have been developed. As described in the following subsections, all methods contribute to finding bad spectra in different ways.

5.11.1 Bad Spectra Recognition from the Power Spectrum

The power spectrum analysis of MS/MS spectra in this work also indicated a criterion that can be used for the identification of bad spectra that are not useful for further study. Detection and removal of bad spectra saves computational time for the spectra interpretation and possible false interpretation results can be avoided. Two types of irregularities that coincide with hard-to-interpret protein MS/MS spectra were observed:

(i) the first power spectrum can exhibit very low amplitudes for low frequencies (the power spectrum is shifted),

(ii) finding the most prominent frequency in the second power spectrum can be ambiguous (several similarly high peaks). In both cases, the cleaning procedures for background removal cannot be straightforwardly applied and, therefore, each mass spectrum is subjected to a routine check during analysis.

With the base frequency derived from the second power spectrum (Figure 5.24), it is possible to compute the position of expected maxima and minima in the first power spectrum (Figure 5.25). The calculated theoretical maxima and minima have to be confirmed by the real data.

To check if the power spectrum is shifted it is necessary to determine whether the real minima and maxima within periods are, on average, closer to the expected positions or closer to the positions with the shift of half a period. If the spectrum is shifted (i.e., if the sum of distances of real maxima and minima from their expected positions is larger than to the positions with a shift of half a period) away from the expected position of minima/maxima,

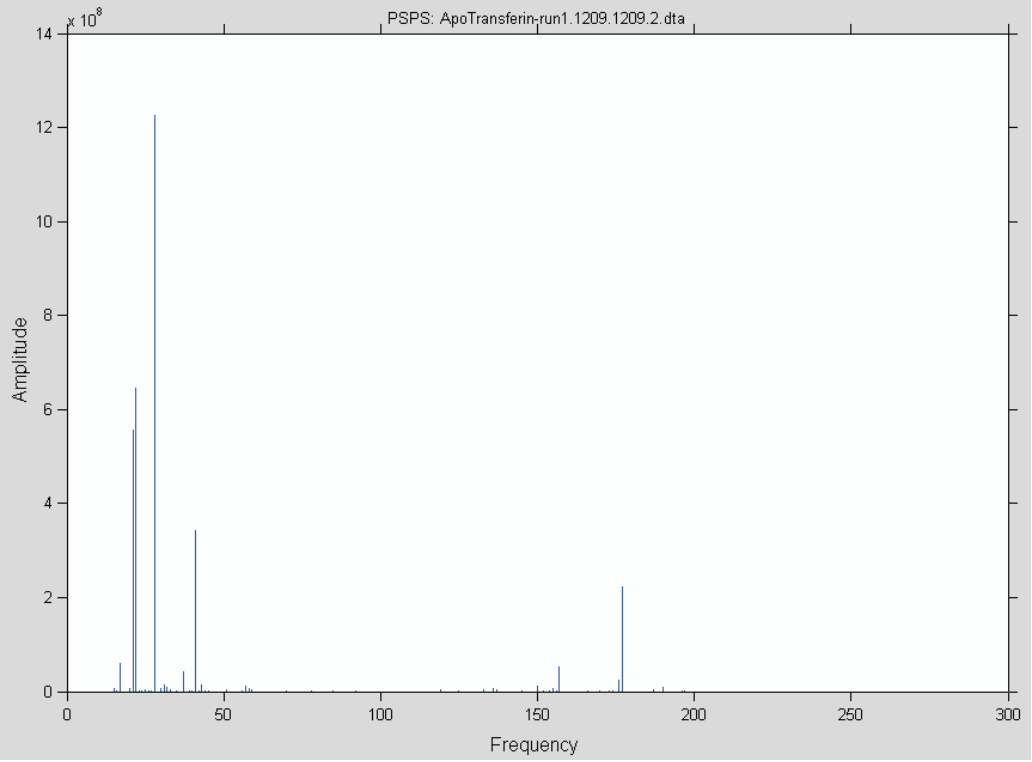


Figure 5.24: Power spectrum shows sometimes non-ambiguous periodicity

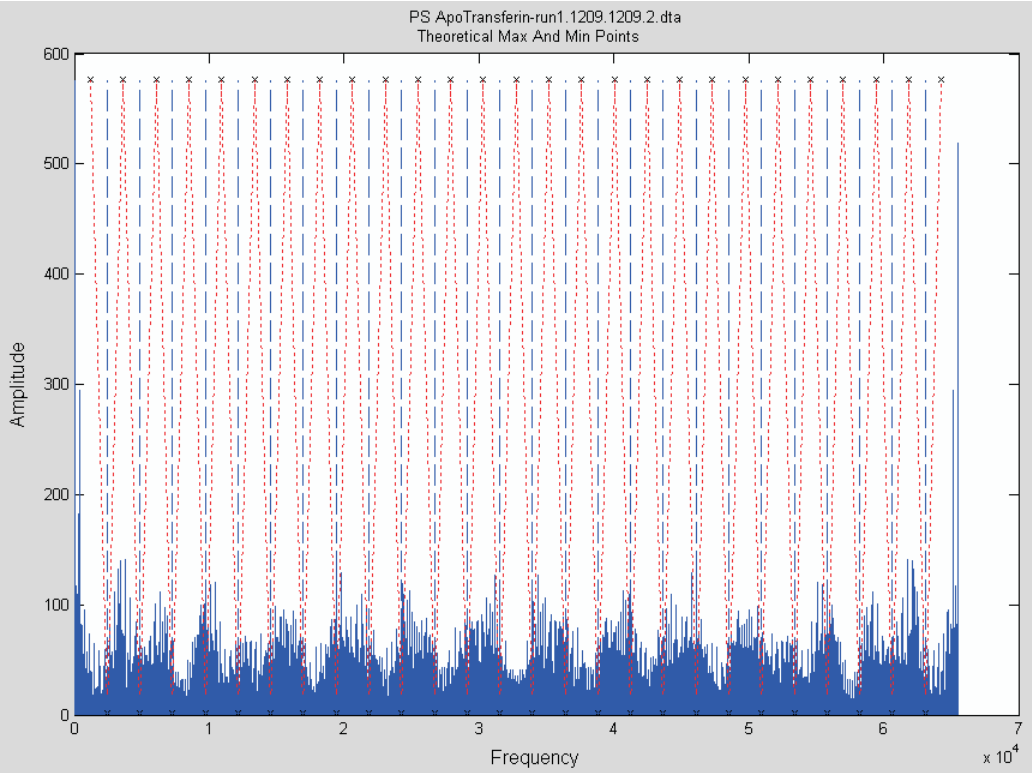


Figure 5.25: Theoretical calculated minima and maxima of the power spectrum considering the highest peak in PSPS as the periodicity of the PS

the procedure for deisotoping is halted. Unfortunately, large shifts in the power spectrum away from expected minima/maxima often indicate bad spectra. For making an appropriate decision, the periodicity of the spectrum is also tested with the coefficient of dispersion, a similarly elementary criterion as the shift. The coefficient of dispersion (C_d) of peak distances in the power spectrum, was calculated as the ratio of the standard deviation of peak distances (s) to the mean value of peak distances ($\bar{X}$).

$$C_d = \frac{s}{\bar{X}} \tag{5.6}$$

A C_d close to zero indicates good coincidence of distances between maxima (and, respectively, minima) of consecutive periods with the expected distance (equal to the period length). Large values of C_d signal distorted periodicity in the power spectrum and a periodicity model appears not applicable. Such spectra are returned to further processing without removal of latent periodic noise. The large C_d values are in some cases the result of a false estimation of the periodicity in the power spectrum. This is often the case if a rigorous periodicity detection algorithm is used. The soft periodicity detection algorithm yields less bad spectra because the main periodicity of the PS spectrum is taken which has the lowest coefficient of dispersion.

The case of quasi-periodic but shifted spectra is more complicated. In such a situation, if the coefficient of dispersion is not larger than 3.3 (an empirically derived threshold), the algorithm predicts that the respective MS/MS spectra cannot be reliably analyzed with interpretation software [30]. As will be shown below, spectra flagged with this criterion are indeed not well interpretable even with database search-based software (i.e., no protein hits are found or only hits with very low reliability).

In rare cases, the suppression of very low frequencies in the PSPS-graph leads to incorrect base frequency determination (fB that is too high) and, consequently, to apparently shifted spectra. These few spectra marked as non-interpretable are false-positively rejected and represent part of the price for automatically cleaning large-scale MS/MS measurements from background with spectral methods as described here.

5.11.2 Bad Spectra Recognition with SNR

Although a certain amount of bad spectra could be detected with the method described above, this method alone is not sufficient. The bad spectra found were really non-interpretable spectra with very low false positive rate. The problem with the method is that only a few percent of all bad spectra could be detected. Three other methods were developed and tested. In this section, signal-to-noise ratio (SNR) is described.

To characterize a spectrum by its SNR, we need to know what is the signal and what is the noise in the spectrum. In some other signal processing problems, the frequency of the noise signals is mostly already known. In the case of MS/MS spectra, this information is not available. To calculate the SNR for MS/MS spectra, the signal and noise fraction had to be defined. This method is based on the peaks intensity in an MS/MS spectrum. High intensity peaks are considered as true peaks, and lower intensity peaks as noise. It is not possible to define an apsolute intensity threshold for all spectra. Instead of defining a threshold, a percentage of high intensity peaks is defined as true signal, and a percentage of low intensity peaks is defined as noise. For a given percentage of noise and peak signals in MS/MS spectrum, two numbers were calculated for each spectrum:

n Number of peaks declared as signal peaks

m Number of peaks declared as noise peaks

Now, the signal-to-noise ratio can be calculated as follows:

$$SNR = 10 \cdot \log_{10} \frac{S_{eff}}{N_{eff}} \; [db] \tag{5.7}$$

where:

$$S_{eff} = \frac{1}{n} \cdot \sum_{i=1}^{n} Intensity_i^2 \tag{5.8}$$

$$N_{eff} = \frac{1}{m} \cdot \sum_{i=1}^{m} Intensity_i^2 \tag{5.9}$$

It was expected that good spectra should have much higher SNR values than non-interpretable spectra. An example of the SNR distribution calculated for 2376 MS/MS spectra of ADH probe (see results chapter 7) is presented in the Figure 5.26.

The next task is to derive parameter values for the SNR threshold, the percentage of the noise in spectrum and the percentage of the signal. Table 4.1 shows results of experiments with BSA data. Different signal percentages were assumed and for each of them SNR was calculated for every spectrum from a set of 2680 BSA spectra. Separately from this calculation, SNR was calculated only for the subset of interpretable files among all files (SNR_{min}). Setting the SNR threshold to SNR_{min}, it was possible to extract 200 to 350 bad spectra from set of 2680 spectra. Compared to the previous method, this method found 20 to 60 new bad spectra. The results are shown in the table 4.1.

Although Mascot could not interpret spectra marked as non-interpretable

Signal [%]	SNRmin [db]	SNR Cut-Off	Bad spectra detected	Mascot could interpret	New bad spectra detected
10	17.2	17.15	220	0	26
20	17.75	17.7	257	0	36
30	18.84	18.8	317	0	52
40	19.18	19.15	337	0	54
50	19.39	19.35	346	0	56
60	19.43	19.4	339	0	53
70	19.43	19.4	324	0	49
80	19.53	19.5	319	0	47
90	19.84	19.82	332	0	49

Table 5.1: Signal-to-noise tests dependent on the given signal percentage. In the first column is listed the given percentage of signal peaks in the spectrum. The second column lists the smallest SNR values of interpretable spectra. The third column lists chosen SNR cut-off values in order to remove all bad spectra with SNR below this value. The fourth column shows how many bad spectra could be detected with this method. The fifth column lists the number of spectra with SNR below the SNR cut-off which could be interpreted by Mascot. The last column contains the number of spectra that could be recognized only with this method and not with the method described in section 5.11.1

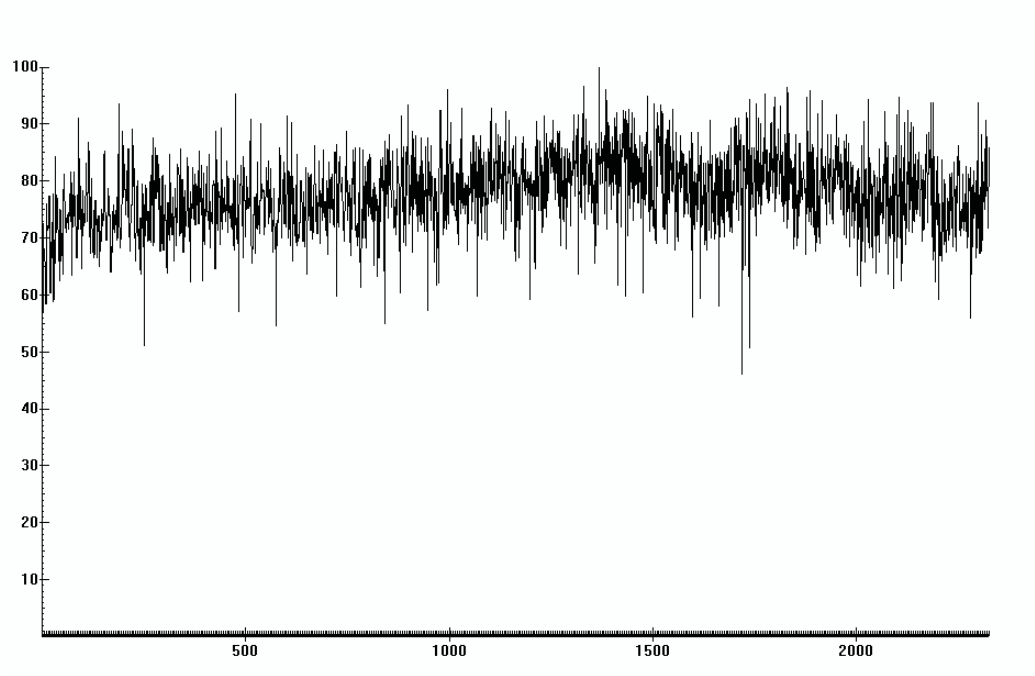


Figure 5.26: Signal-to-noise ratio distribution found in several thousand MS/MS spectra

spectra with this method, the number of detected non-interpretable spectra is still small.

5.11.3 Bad Spectra Detected by Signal Entropy

As possible criteria for bad spectra detection, the weighted and non-weighted entropy value was calculated for every MS/MS spectrum.
Weighted entropy:

$$E_w = -\sum_{i=1}^{n} P_i \ln(P_i) \tag{5.10}$$

Non-weighted entropy:

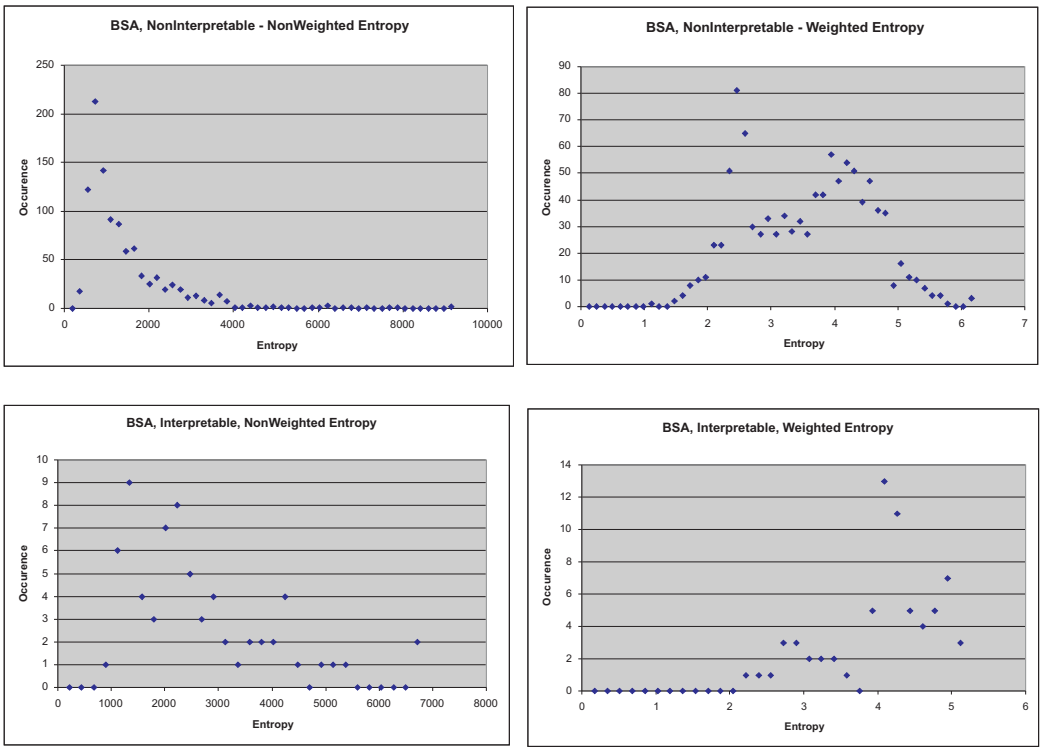


Figure 5.27: Weighted and non-weighted entropy of interpretable and non-interpretable spectra

$$E_w = -\sum_{i=1}^{n} \ln(P_i) \tag{5.11}$$

With P_i defined as:

$$P_i = \frac{Intensity_i}{\sum_{i=1}^{n} Intensity_i} \tag{5.12}$$

The results of this experiment are presented in Figure 5.27. The figure shows that some spectra could be successfully marked as bad spectra if an appropriate threshold was chosen. Several data sets with thousands of spectra were analysed but no uniform rule could be derived.

Chapter 6

Implementation

6.1 Computer Program "MS Cleaner"

The algorithms and methods developed during this thesis were implemented in a computer program called "MS Cleaner". The program is developed with Microsoft Visual Studio 2005 integrated developement enviroment in C++ programming language. It runs on a single Windows PC as well as on a Linux cluster. A copy of the program is available for free download at: http://mendel.imp.ac.at/mass-spectrometry/.

The Windows version is a multithreaded application consisting of 3 components (Figure 6.1):

(i) User interface,

(ii) Main program,

(iii) Library of MS handling functions (MSManager),

(iv) Library of DSP functions (DSPLib).

The user interface and main program were implemented in one exe-

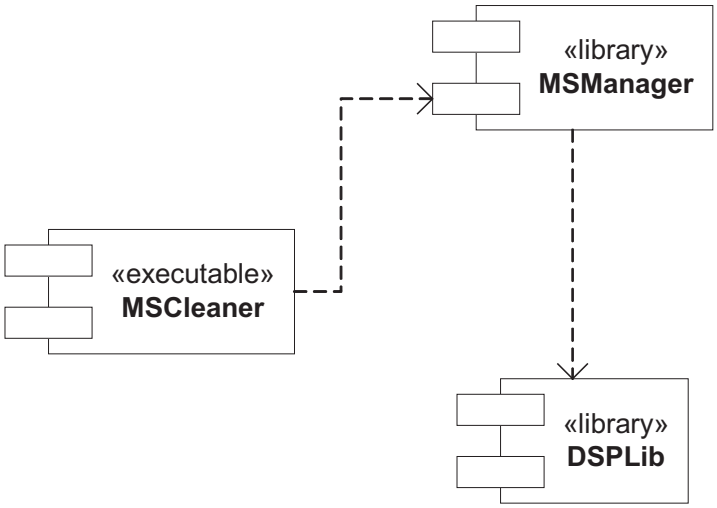


Figure 6.1: MSCleaner component diagram

cutable file, while the libraries were developed as separate dynamic linked libraries. The user interface was implemented as an MFC (Microsoft Foundation Classes) Windows application.

The two libraries were developed as COM (Component Object Model) objects.

Interfaces and functions of the "MSManager" COM object are:

1. *IImportMSMS* interface

   - ImportSequestDTA(const char* pchPathName, CDta& dta)

   - This function imports MS/MS spectra in *dta* format (see subsection 6.1.1).

   - ImportNextDtaFromMGF(CStdioFile& fPos, CDta& dtaOut)

   - The function imports single *dta* files from an *mgf* file.

   - ImportFinniganASC(char* pchPathName)

   - The function imports MS/MS spectra from Finnigan *raw* files.

2. *IConvertMSMS* interface

   - RawToMgf(CString strRAW, CString strMGF, CXRawFileEx* pRawFile, Raw2MgfParams* pRP, CEvent* pKillEvent, CWnd* pWnd)
   - The function converts Finnigan *raw* files into *mgf* files.
   - RawToDta(CString strRAW, CString strExportPath, CXRawFileEx* pRawFile, Raw2MgfParams* pRP, std::list<std::string>& lstrDtams, std::list<std::string>& lstrDtams2, std::list<std::string>& lstrDta, CEvent* pKillEvent, CWnd* pWnd)
   - The function converts *mgf* files into *dta* files.
   - DtaToMgf(const std::list<std::string>* plstDtaFileNames, std::string strExportPath, CEvent* pKillEvent, CWnd* pWnd)
   - The function converts *dta* files into *mgf* files.
   - MgfToDta(std::string strMGF, std::string strExportPath, CEvent* pKillEvent, CWnd* pWnd)
   - The function converts *mgf* files into *dta* files.

3. *IExportMSMS* interface

   - ExportDTA(const char* pchPathName, const CDta* pDta, int nCommas = 2)
   - The function exports cleaned MS/MS spectra into *dta* files.
   - ExportDTAintoMGF(CStdioFile& fMGF, const CDta* pDta, int nCommas = 2, std::string strMGF = “”)

- The function exports cleaned MS/MS spectra into *mgf* files as a collection of *dta* files.

4. *IEditMSMS* interface

   - MakeEqDist(double dDistance, CDta& dtaToDense)
   - This is an implementation of the algorithm "Make Equidistant Spectrum" (section 5.4).
   - DenseSpectrum(int nChargeToCheck, CDta* pDtaDensed)
   - This is an implementation of the algorithm "Dense Spectrum" (section 5.6).
   - MergePeaks(CDta* pDta, double dLowestDistance)
   - This is an implementation of the algorithm "Merge Peaks" (section 5.3).
   - DeconvoluteMultiCharge(CDta* pDtaDensed,
     const std::vector<PeakIter>* pvPeaksMultiCharged,
     const CIntensityDistrib* pID, bool bCreateLog,
     CStdioFile* pfLog, std::vector<Peak>* pvPeaksToAdd)
   - This is an implementation of the algorithm "Deconvolute Spectrum" (section 5.7).

5. *ICleanMSMS* interface

   - CleanSpectra(const CleanSpecParams* params)
   - This is a generic function for cleaning of MS/MS spectra in all three formats.
   - CleanSequestDTA(const CleanSpecParams* params)

- This is a function for cleaning of MS/MS spectra in *dta* format.

- CleanMGF(const CleanSpecParams* params)

- This is a function for cleaning of MS/MS spectra in *mgf* format.

- CleanRAW(const CleanSpecParams* params)

- This is a function for cleaning of MS/MS spectra in Finnigan *raw* format.

6. *IAnalyzeMSMS* interface

   - FindPeakClusters(const CDta* pDta, CDta* pDtaDensed, int nChargeOfPeakCluster, CIntensityDistrib* pID, double dMinR, std::vector<PeakIter>* pvFoundPeaksMultiCharged)

   - The function scans for isotope peaks clusters according to the charge state supplied by the parameter "nChargeOfPeakCluster".

   - CheckSeqLadder(const CDta* pDta, int nLeastSeqTagNumber, double dMassTolerance, int nSLIntPercentage, bool& bSeqTagsFound)

   - The function returns "true" if a sequence ladder of length "nLeastSeqTagNumber" was found (section 5.2).

   - CheckSNR(const CDta* pDta, double& dSNR)

   - This function returns signal-to-noise ratio for defined percentage of signal and noise peaks (section 5.11.2).

Interfaces and functions of the "DSPLib" COM object are:

1. *IDSPLib* interface

   - Filter(const dVect* pvdB, const dVect* pvdA, const dVect* pvdX, dVect* pvdZi, dVect* pvdY)

- This function filters data with an infinite impulse response (IIR) or finite impulse response (FIR) filter.

- Convolute(dVect* pvdA, dVect* pvdB, dVect* pvdC)

- This function performs convolution and polynomial multiplication.

- Deconvolute(dVect* pvdA, dVect* pvdB, dVect* pvdQ, dVect* pvdR)

- This function performs deconvolution and polynomial division.

- DFour(std::vector<double>* pvdData, unsigned long nn, std::vector<std::complex<double>> * pvcFFT)

- This function performs discrete Fourier transform.

- DIFour(std::vector<double>* pvdData, unsigned long nn, std::vector<std::complex<double>>* pvcIFFT);

- This function performs inverse discrete Fourier transform.

- PowerSpectrum(std::vector<std::complex<double>>* pvcFT, dVect* pvdPS, double& dPSMeanOut)

- This function calculates power spectrum of a given signal.

- MedianFilter(const dVect* pvdPS, unsigned long nSignalSize, unsigned long nFilterSize, dVect* pvdPSMedFiltered, double* pdMean)

- The function implements the algorithm “Median Filter” described in section 5.8.

- EigenValue(double* pdNxNMatrix, int nN, std::vector<std::complex<double>>* pvcEigenValue)

- The function returns a vector of the eigenvalues of a given matrix.

- Roots(std::vector<double>* pvdCoeff, std::vector<std::complex<double>>* pvcRoots)

  - This function returns a column vector whose elements are the roots of a given polynomial.

- Poly(std::vector<std::complex<double>>* pvcCoeff, std::vector<std::complex<double>>* pvcPoly)

  - This function returns a row vector whose elements are the coefficients of the polynomial whose roots are the elements of a given vector.

- PolyStab(std::vector<double>* pvdCoeff, std::vector<double>* pvdPolyStab)

  - The function finds the roots of the polynomial and maps those roots found outside the unit circle to the inside of the unit circle.

- Toeplitz(std::vector<double>* pvdC, std::vector<double>* pvdR, std::vector<double>* pvdT)

  - The function returns a nonsymmetric Toeplitz matrix T having "pvdC" as its first column and "pvdR" as its first row.

- FreqResponse(const std::vector<double>* pvdB, const std::vector<double>* pvdA, long int lnSize, std::vector<std::complex<double>>* pvcH)

  - The function returns the complex frequency response (Laplace transform) of an analog filter.

- InverseMatrix(std::vector<double>* pvdA, int nNxNMatrixOrder, std::vector<double>* pvdInvA)

  - The function returns the inverse of a given square matrix.

- MatrixMultiplication(std::vector<double>* pvdA, int nOrderA, std::vector<double>* pvdB, int nOrderB, std::vector<double>* pvdResult)

  - The function multiplicates two given matrices.

- MatrixDivision(std::vector<double>* pvdA, int nOrderA, std::vector<double>* pvdB, int nOrderB, std::vector<double>* pvdResult)

  - The function divides two given matrices.

- Yulewalk(int nOrderSize, std::vector<double>* pvdFrequency, std::vector<int>* pvnMagnitude, std::vector<double>* pvdB, std::vector<double>* pvdA)

  - This function designs recursive IIR digital filters using a least-squares fit to a specified frequency response.

- LowPassFilter(CLowPassFilter::BandStopFreq w, const std::vector<double>* pdIn, std::vector<double>* pdOut)

  - The function returns a lowpass filter with a desired cutoff frequency "w" in normalized frequency (Nyquist frequency = 1 Hz).

The Linux cluster version does not have user interface. The main program and all functions are compiled in one executable file.

6.1.1 Input Data

The spectra cleaning and pre-processing developed in this work is a practice-oriented solution which allow large scale analysis of real data from MS labs suitable to be incorporated in the existing protein analysis before the MS

data are sent to interpretation software. Data supported for processing are in the following formats:

(i) dta files,

(ii) mgf files and

(iii) raw files.

Dta files are text files describing a single MS/MS spectrum. Dta files contain precursor mass (floating comma value) and intensity (integer value) in the first row. All other rows in the file contain m/z and intensity values (as floating comma values).

Mgf files are text files with multiple MS/MS spectra, mostly all dta files from an MS analysis merged together.

Raw files are binary data in Thermo Finnigan format containing all MS and MS/MS spectra as well as all relevant parameters used for the experiment.

MS Cleaner supports input and conversion between these three file formats. As output only dta and mgf files are supported.

6.1.2 User Interface

User interface of the program "MS Cleaner" is implemented as *Dialog* based Windows application. Dialog box and options tabs are shown in figures 6.2, 6.3, 6.4.

Parameters and control elements of the program are:

- *Data files* (string) - This list box displays the input data files to be cleaned.

- *Export cleaned spectra into directory* (string) - This edit box displays the directory into which the cleaned spectra will be saved.
- *Options* (boolean) - Clicking this button opens and closes a form allowing user to change the basic, internal and spectra extraction parameters.
- *Default* - A button to reset all parameters to their default values.
- *Help* - Clicking this button invokes a *Help* document.
- *Start* - This button starts the cleaning procedure.
- *Close* - This button closes the dialog box and quits the program.
- *Create log file* (boolean) - Selecting this check box will result in generation of log file, which for each spectrum records the results of the cleaning procedure.
- *Deconvolution of multi-charged peaks* (boolean) - If selected, the program detects multi-charged isotope peak clusters according to the chosen *Deconvolution threshold*.
- *Deisotoping* (boolean) - If selected, the program detects singly-charged isotope peak clusters according to the chosen *Deisotoping threshold*.
- *Random noise removal* (boolean) - If selected, the program removes random noise from the spectrum according to the *Random noise threshold*.
- *Check for sequence ladder* (boolean) - When this option is selected, the program inspects each MS/MS spectrum for a series of peak spacing

with m/z values corresponding to amino acids mass (with mass tolerance value of the *Mass tolerance* parameter). The user can define the length of the sequence ladder in the accompanying-box (values in the range 3 to 5 are recommended).

- *Do only data conversion* (boolean) - Selecting this check box will result in perorming only a data files conversion from raw files without starting the cleaning procedure.

- *Merge output mgf files* (boolean) - Selecting this check box will result in merging all output mgf files into a single file.

- *Determines bad spectra handling* (array of bool values) - Selecting the first option will leave spectra designated as "bad" in the "cleaned" output directory. Using this option doesn't separate bad spectra from remaining files. Selecting the second option creates two sets of output file(s):cleaned files (minus the bad spectra), and bad spectra files. Selecting the third option deletes the bad spectra from the output files.

- *Merge peaks* (boolean) - If selected, the program merges together peaks that are closer than the *Least Peaks Spacing* value.

- *Least peaks spacing* (double) - Specifies the least allowed spacing between peaks for the *Merge peaks* algorithm of the program.

- *Median filter* (integer) - Specifies how many signals are taken for median filtering of the power spectrum.

- *Deisotoping filter width* (double) - Specifies the deisotoping filter width.

- *Max PS frequency* (integer) - Specifies the maximum allowed frequency of the power spectrum.

- *Sequence ladder length* (integer) - Specifies the sequence ladder length for the *Check sequence ladder* algorithm.

- *Sequence ladder intensity* (integer from 0 to 100) - Specifies a percentage of the most intensity peaks to include in the *Check sequence ladder* algorithm.

- *Rigorous detection of bad spectra* (bool) - This mode is used to find the highest possible number of bad spectra and reduce the amount of data as much as possible. Although some spectra are bad, they can still be interpreted by a database search program. With this option, removal of bad spectra can lead to a lower sequence coverage, but a higher confidence in the interpretation.

- *Soft detection of bad spectra* (bool) - Using this option, fewer bad spectra will be identified. This mode is used if a high sequence coverage is more important than data reduction.

- *Deisotoping threshold* (integer from 0 to 100) - The deisotoping procedure reduces the intensity of the potential isotope peaks. If the decrease in intensity is greater than this threshold value, the intensity of the peak will be set to zero.

- *Deconvolution threshold* (integer from 0 to 100) - The deconvolution procedure detects multi-charged peaks if the correlation coefficient is higher than this value.

- *Random noise filter* (double) - This is the stop frequency for the low pass filter used for random noise removal.

- *Random noise threshold* (integer from 0 to 100) - This filter reduces the

intensity of random noise peaks. If the decrease in intensity is greater than this value, the intensity is set to zero.

- *Mass tolerance* (double) - Mass tolerance taken into account during the sequence tag determination.
- *FT MS* (bool) - Indicates that input data were produced by an FT MS instrument.
- *LTQ* (bool) - Indicates that input data were produced by an LTQ MS instrument.
- *LCQ* (bool) - Indicates that input data were produced by an LCQ MS instrument.
- *MW range from, to* (two integer values) - Specifies m/z range of precursor ions.
- *Threshold* (absolute, relative) - Specifies an intensity threshold of precursor ions.
- *Minimum ion count* (integer) - Specifies a min number of peaks that an MS/MS spectrum needs to have.
- *Separate MS2 from MS3* (boolean) - This option is meaningful only if a "raw" file contains MS3 spectra as well.
- *Convert '.raw' files to '.dta' ('.mgf'))* (boolean) - Specifies the output file format.

6.2 Other Tools Developed

Two other tools were developed to analyze the MS/MS spectra.

6.2.1 Tool for Creating Theoretical Fragment Ions from Protein Sequences "DigestIt"

In many cases, the mass spectrometry is not used to identify protein in the probe but to examine the characteristics of an already identified protein. The best example is the analysis of protein post-translational modifications. If the protein is known, the first question of importance would be: what spectra are expected to be found in the set of all MS/MS spectra. To avoid manual creation of fragment ions, a tool was developed to automatically perform that task. As input, the program needs a protein sequence in text format, known modification on the protein and the enzyme used for the digestion. Predefined lists of modifications and enzymes can be extended by creating or modifying new modifications or enzymes. The tool is especially useful in the case of analysing post-translational modifications on proteins where sequences cannot be identified by database search programs.

6.2.2 MS Fragmentation Viewer

Low energy collision induced dissociation causes fragmentation mainly across the peptide bond. The peptide bond energy is different between amino acids and the difference should be represented in MS/MS spectra. Studying the distribution of fragment ions in the spectrum is the first step toward creating reliable scoring function for any MS/MS spectra interpretation software. For this purpose, a fragmentation viewer was developed displaying intensity distribution of different ion types in MS/MS experiments. The program takes

interpretation results from Mascot and calculates different fragment ions and counts their occurrence in the interpreted data (Figure 6.6).

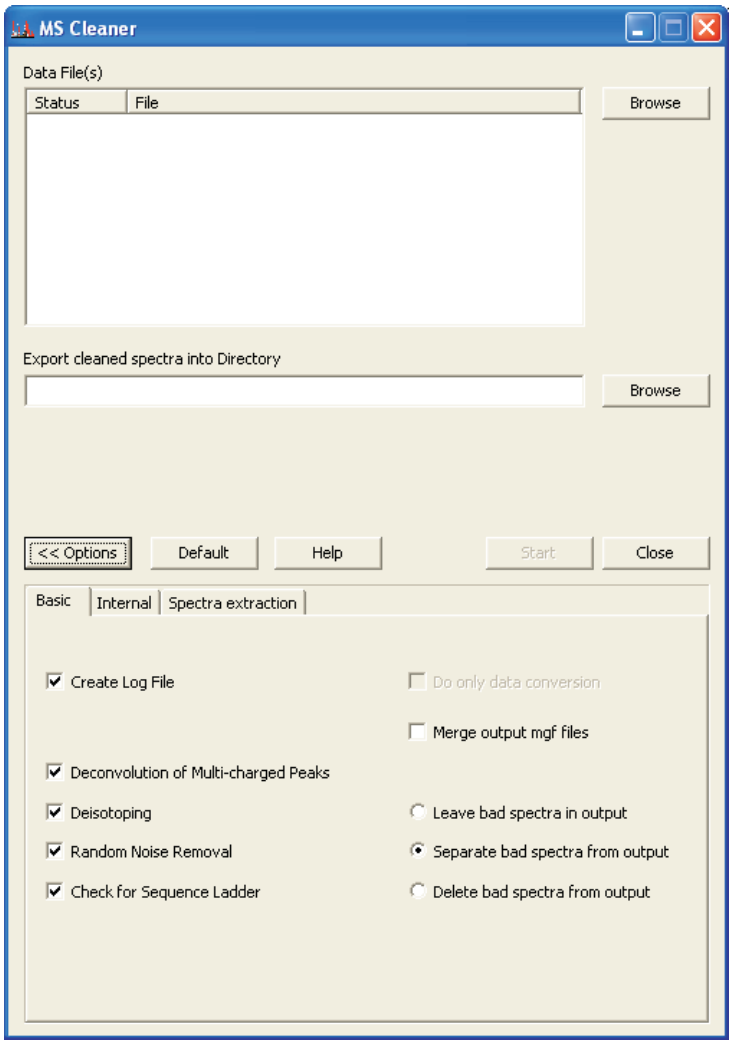


Figure 6.2: User interface of "MS Cleaner" with "Basic Options" tab selected

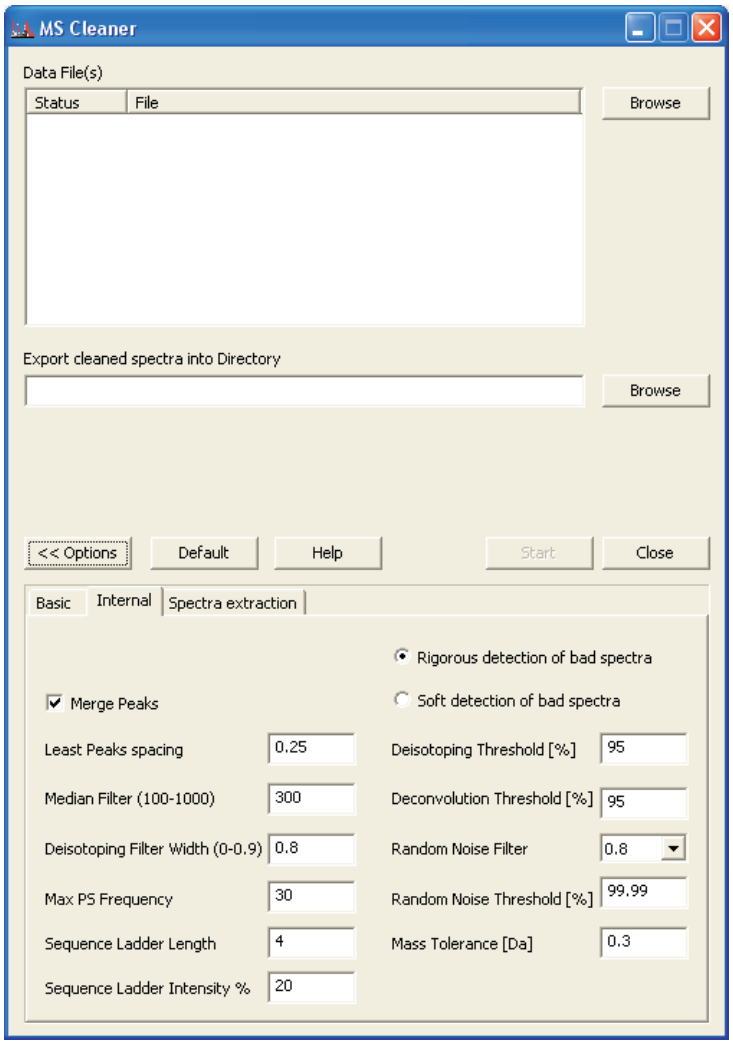


Figure 6.3: User interface of "MS Cleaner" with "Internal Options" tab selected

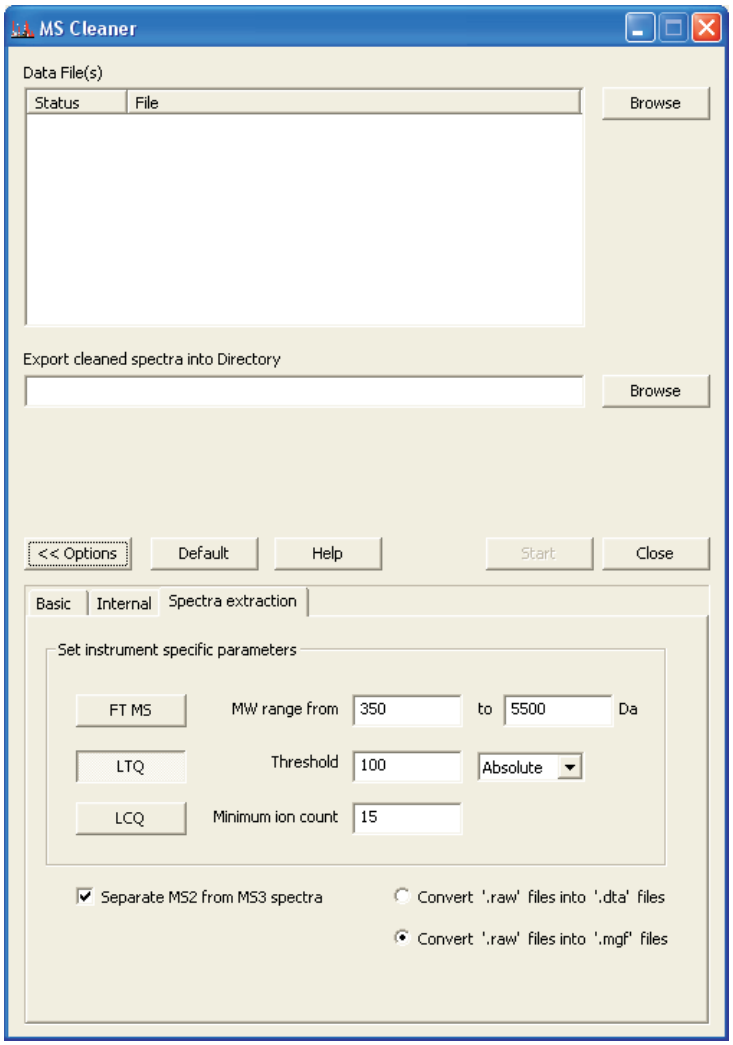


Figure 6.4: User interface of "MS Cleaner" with "Spectra Extraction Options" tab selected

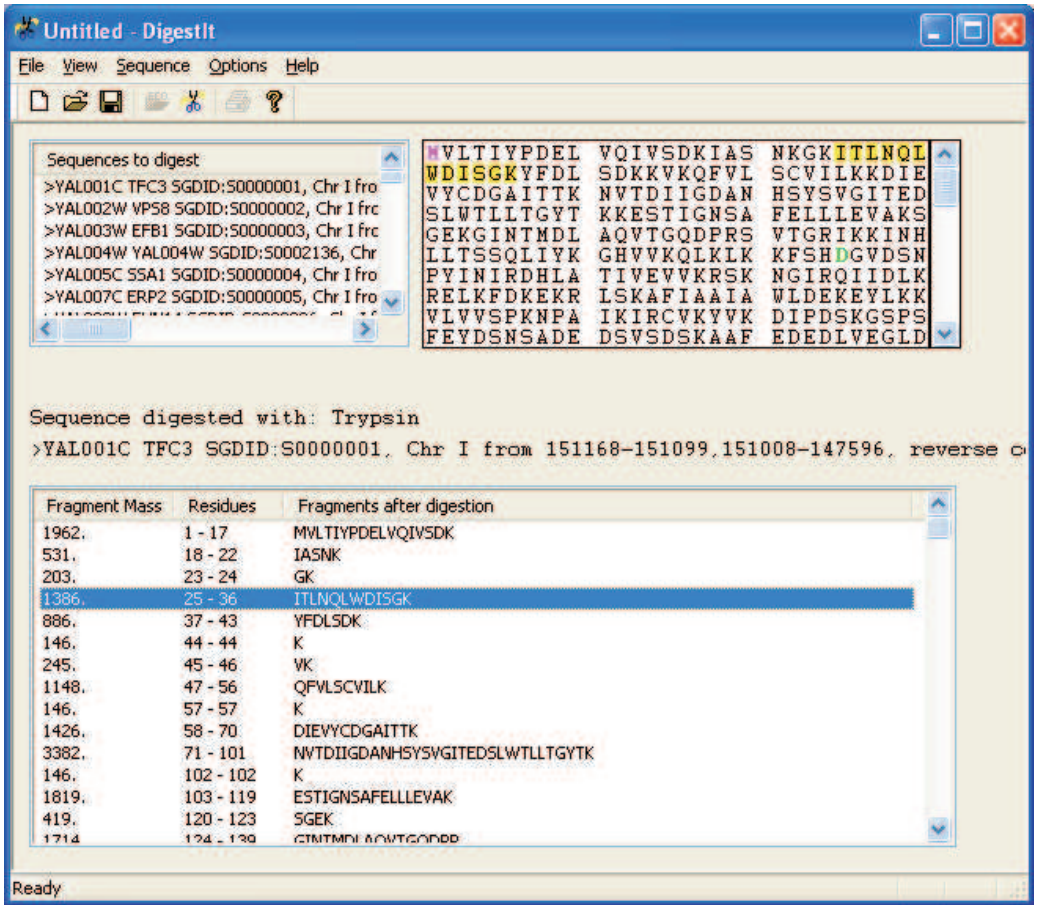


Figure 6.5: Computer program "Digest It" developed as a tool to examine expected precursor ions in MS and MS/MS spectra

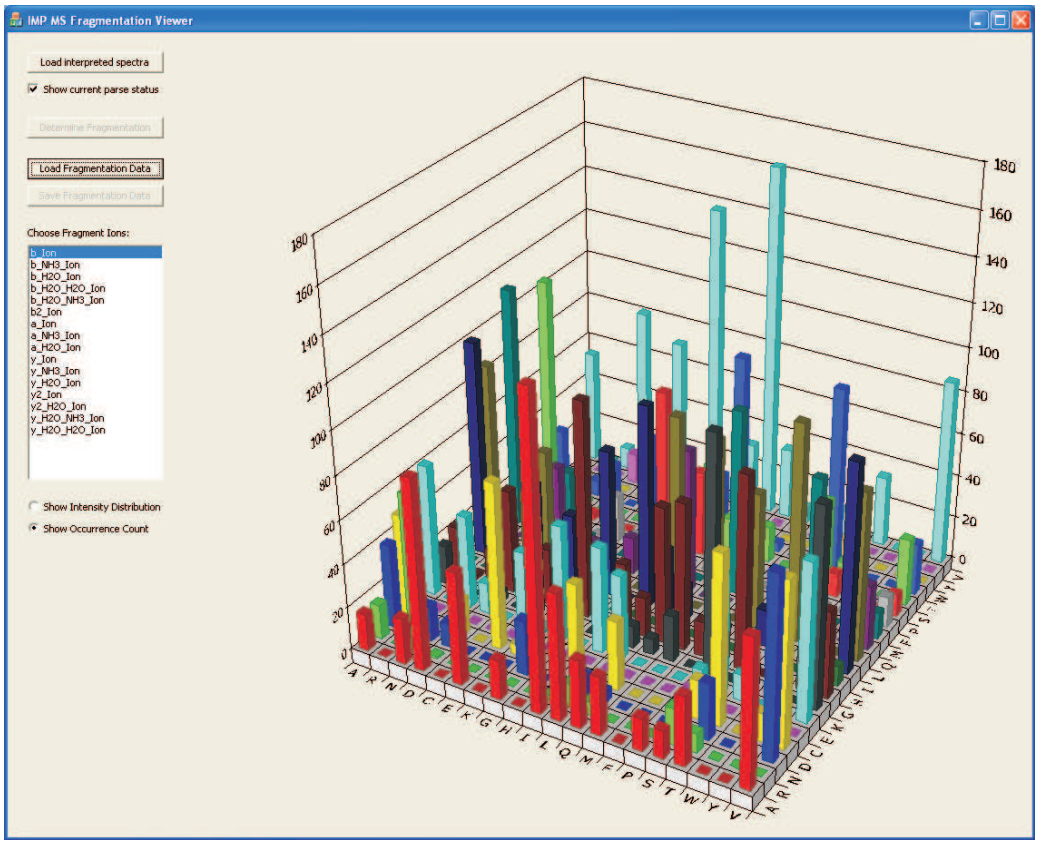


Figure 6.6: Computer program "MS fragmentation viewer" developed as a tool to examine the distribution of fragment ions in the spectrum

Chapter 7

Experimental Results

7.1 General Considerations for Testing Procedures for Background Removal in Tandem Mass Spectra

In the ideal world, background removal algorithms would be parameterized and tested against a large library of MS/MS spectra where the different types of all noise (e.g., multiply charged peaks, isotope clusters, random noise, etc.) are explicitly annotated in electronically readable form and the rates of true- and false-positive detection of various noise types can be directly computed. Unfortunately, such a library was not available during this research effort and its creation is beyond the scope of this work. The background removal algorithms were validated implicitly. The automated interpretation of MS/MS spectra with MASCOT has become a virtual standard in proteomics laboratories; therefore, the MASCOT-generated interpretations both for the original MS/MS spectra and the spectrum versions after the application of our background removal procedure were compared. Discrepancies between

both interpretations can be automatically detected in large-scale tests of real datasets and summarized by computer programs. The parameters described above have been selected to achieve a minimum of cases of accidental removal of peaks that are relevant for interpretation by MASCOT in large-scale tests.

Two sets of large scale tests have been performed during this work. The first test was aimed to clean interpretable spectra; to improve sequence coverage by increasing quality of peptide spectra that could not be interpreted before cleaning; to increase the score of peptide spectra by removing background peaks and transforming heavy isotope peaks (singly and multiply charged) into singly charged monoisotopic peaks.

The second set of large scale tests was performed in order to find as many as possible bad spectra reducing on that way the total computational time of the following interpretation step.

7.2 Tests on Improvement of the Quality of Interpretable Spectra

Results of background removal in MS/MS spectra obtained with 100 fmol BSA, ADH and TRF. To test the MS Cleaner in practical large-scale applications, MS/MS spectra from protein samples with known composition were used. Such spectra of well known proteins such as BSA, ADH or TRF are regularly produced for the purpose of quality control of MS instrumentation with low concentrations (for example 100 fmol). Original and cleaned spectra as well as supplementary tables that show changes of scores of leading peptide hits are available at the associated WWW-site (http://mendel.imp.ac.at/mass-spectrometry/).

The respective dta-files were merged to generate a single mgf-file (Mascot

generic format) using the merge.pl program (Matrix Science). This original mgf-file was then processed using the MS Cleaner program, using the default internal parameters, generating two new mgf-files with cleaned and bad spectra respectively. All three mgf-files were used to perform Mascot MS/MS Ions Searches (Matrix Science). In the case of BSA, ADH, and TRF, the non-redundant protein sequence database was used (as of 15th of December, 2005). In the case of the condensin sample, the identification of posttranslational phosphorylations was the original task. Therefore, the search was initially performed against a small curated protein database (146 sequences; 68753 residues), which includes components of the condensin, cohesin, and kinetochore complexes, as well as some common contaminants and trypsin, in the case of the condensin sample. Additionally, searches against all human as well as against all proteins in the non-redundant database were carried out. It should be noted that the Mascot score for recovering the original proteins tend to be the higher, the smaller the database due to reduced sequence background; thus, the search with the small database of 146 sequences is the more stringent condition compared with searches in the non-redundant database. The Mascot search parameters were the same in all runs (enzyme: trypsin; fixed modifications: carbamidomethyl (Cys); variable modifications: oxidation (Met); peptide charges: 1+, 2+ and 3+; mass values: monoisotopic; protein mass: unrestricted; peptide mass tolerance: 2 Da; fragment mass tolerance: 0.8 Da; max. missed cleavages: 1). The Mascot search results output html-file was formatted with standard scoring, a significance threshold of p¡0.05, and an ion score cut-off for each peptide of 30.

The results of applying the background removal procedure are summarized in Table 7.2, and Table 7.2. First, it is evident that protein hits are found from the cleaned MS/MS spectra with considerably increased scores.

This is evident for the total protein score (between 10% and 15%, see Table 7.2). Scores improve for the majority of all leading peptide hits (about 70%, see Table 7.2). A decrease is observed for about 10% of the cases but did not affect the interpretation except of one case (see below). In general, the likelihood of retrieving the sample protein and the sequence coverage improve (see Table 7.2). This conclusion is in line with the logics of MS/MS spectra interpretation schemes such as Mascot: The MS Cleaner-based background removal decreases the number of peaks considerably. Therefore, the number of alternative (including false-positively hit) protein sequences that might fit a given spectrum reduces and the scores of the top hits against the alternatives naturally improve.

MS/MS spectra considered non-interpretable by our procedure are indeed bad spectra. In only one out of 626 cases was the original protein recovered by Mascot. Here, Mascot assigned a score of 64 (see Table 7.2 and also data and figures at mendel.imp.ac.at/mass-spectrometry/falsepositive-partA.html). Visual inspection of the spectrum revealed almost no significant peaks above background. This single artifact of rejection by MS Cleaner is a result of the suppression of low frequencies in the PSPS-graph and would disappear with a slightly reduced threshold. In contrast, there are a considerable number of spectra (about 10%) that become interpretable for Mascot only after background removal with our procedures (5 for BSA, 1 for ADH, 8 for TRF, see Table 7.2). An example is shown in Figure 7.1. Figure 7.1-A represents an original MS/MS spectrum of 100 fmol BSA (abscissa: m/z in Da, ordinate: relative intensity; totally 373 peaks). Background peaks that have been removed by MS Cleaner are shown in blue (83), other peaks are shown in red (290). Figure 7.1-B is Mascot interpretation of the cleaned spectrum (as peptide sequence LVTDLTK). The spectrum is shown with as-

Search	dta-files	Score	Match	Cov (%)
BSA				
Raw spectra	2679	1844	65	51
Cleaned spectra	2484	2094	70	56
Bad spectra	195	195	n/a	n/a
Yeast ADH				
Raw spectra	2325	536	24	29
Cleaned spectra	2060	594	25	29
Bad spectra	265	n/a	n/a	n/a
Human TRF				
Raw spectra	2608	1643	61	41
Cleaned spectra	2442	1846	65	44
Bad spectra	166	64	1	2

Table 7.1: Influence of background removal on the recovery of BSA, ADH, and TRF in MS/MS spectra of 100 fmol test samples. The MS/MS spectra were interpreted with MASCOT directly ("raw spectra") and after processing with the background removal procedure ("cleaned spectra") described in this article. The "score" is the MASCOT score from all successful searches; "match" is the number of searches that recover the peptides from the protein used. "cov (%)" reports the sequence coverage. The line "bad spectra" reports the number of files that are considered not "interpretable" by the criterion described in the text (n/a - not applicable). Only in one case could MASCOT recognize a peptide from the original protein in a bad spectrum that is visually also of low quality.

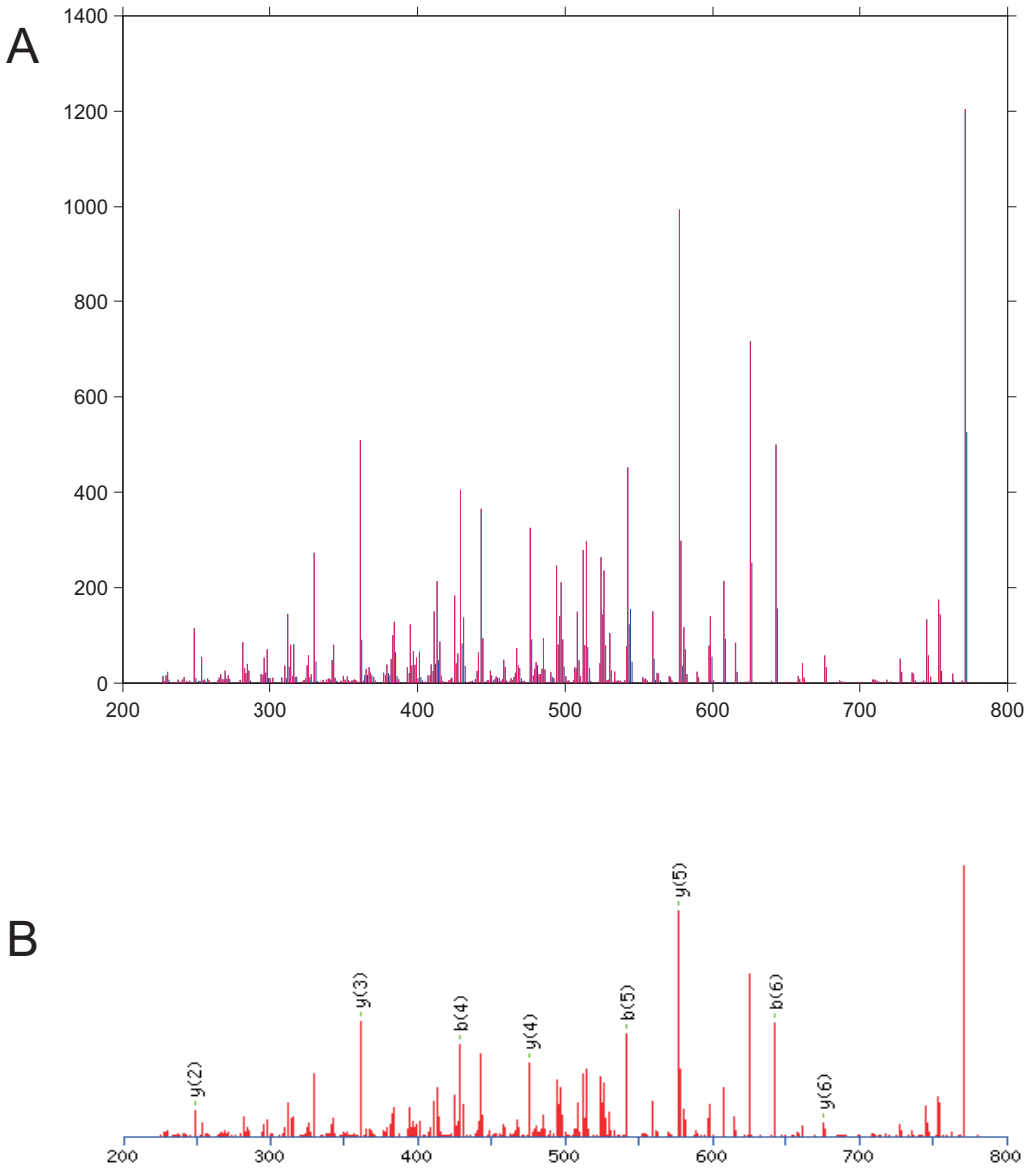
	BSA	ADH	TRF
Total peptide hits	70	25	68
Scores increased	47	18	48
Scores unchanged	5	4	3
Scores decreased	13	2	6
Hits only after cleaning	5	1	8
Hits lost after cleaning	0	0	3

Table 7.2: Changes of scores of leading peptides in MASCOT searches as a result of background cleaning (summary digest of Supplementary tables at the website http://mendel.imp.ac.at/mass-spectrometry/)

signment of b- and y-ions and the table representing the sequence ladder. Out of the 373 peaks in the spectrum, 83 are recognized as background and are removed. As a result, Mascot was no longer confused and was able to assign a full y-series and many b-ions. Although all procedures described in this work are essential for various aspects of background reduction, they contribute differently from the quantitative point of view.

As can be seen from the data in Table 7.2.2, the spectral-analytic criteria (removal of latent periodic and high-frequency noise) are most efficient in reducing the background since their share among the removed peaks is above 90%. In the BSA, ADH and TRF applications, about 15% of all peaks in the original spectra get removed by our program and the file storage requirement is reduced by the same amount. The computational performance of MS Cleaner was tested on a stand-alone PC (Intel(R) Pentium(R) Processor, 2.4GHz, 1GB RAM, Windows XP operating system).

For the BSA case, 2679 dta-files were cleaned in 4:52 min (0.11 sec per spectrum). The Mascot time on the same machine reduced from 64 min (for the untreated data) to 57 min (cleaned files). The respective numbers for ADH (2325 files) and TRF (2608 files) are 5:36 (0.14 sec per file), 75, 64 and



#	b	Seq.	y	#
1	114.09134	L		7
2	213.15975	V	676.38757	6
3	314.20743	T	577.31916	5
4	429.23437	D	476.27148	4
5	542.31843	L	361.24454	3
6	643.36611	T	248.16048	2
7		K	147.11280	1

Figure 7.1: Example of a spectrum that was only interpretable after background removal

4:15 (0.10 sec per file), 58, 50 (all values in minutes). Thus, savings of computational costs are considerable under the condition of increased reliability of spectrum interpretation.

7.2.1 Detailed Analysis of MS Cleaner's Removal of Multiply Charged Peaks in the dta-Files of the BSA Set

It was interesting to check whether the multiply charged peaks assigned by Mascot are detected by the program MS Cleaner. After having manually analyzed the whole BSA dataset, we found only two peaks interpreted as doubly charged by Mascot that had also a remnant isotope cluster (in the dta-file 369.369.2, see supplementary data at http://mendel.imp.ac.at/mass-spectrometry/beforeafterBSA.htm). For this spectrum, MS Cleaner revealed 7 doubly charged clusters. Two of them (at m/z=315.70 and 320.30) include the two doubly charged peaks found by Mascot. The other five are composed of noise peaks. It should be noted that spectral procedures (as a rule, the algorithm for high frequency noise removal) mark many low intensity peak clusters (comparable with the five latter ones) as noise, too. As discussed above, MS/MS measurement accuracy and scanning speed on many instruments prevent the detection of isotope clusters in many cases. The algorithm for detecting multiply charged clusters will work the better, the more accurate the spectra are recorded (as in the new generation of Fourier-Transformation instruments) and the more complete isotope clusters are represented in the data.

7.2.2 Application of the Background Removal to the Condensin Dataset

It should be noted that, in the latter example of BSA, ADH and TRF, low concentrations of proteins are intentionally applied to achieve limiting cases of mass spectra. The analysis of the condensin complex mass spectra is a more biologically relevant application. For this purpose, condensin complexes from cultured human HeLa cells were purified and analyzed. Human cells contain two distinct condensin complexes, called condensin I and condensin II, which bind chromosomes specifically in mitosis and contribute to their condensation and structural integrity [68, 69, 70, 71]. Both complexes are hetero-oligomers composed of five subunits. Two ATPase subunits of the structural maintenance of chromosome (SMC) family, called Smc2 and Smc4, are shared between condensin I and condensin II. In addition, each complex contains a set of distinct non-SMC subunits, called kleisin-γ [70], CAP-G and CAP-D2 in the case of condensin I, and kleisin-β [70], CAP-G2 and CAP-D3 in the case of condensin II. Both complexes were immunopurified simultaneously using antibodies to their common Smc2 subunit and analyzed the resulting sample both by SDS-PAGE and silver staining (Figure 7.2) and by in-solution digest followed by LC-MS/MS. Silver staining revealed bands that correspond to Smc2, Smc4 and to all six non-SMC subunits that are present in condensin I and condensin II. The MS/MS spectra were processed using the MS Cleaner.

All three datasets, the original, the cleaned and the bad spectra, were used to perform a Mascot MS/MS Ions Searches against a small and curated protein database as well as against the non-redundant protein database (all proteins and all human proteins).

A summary of the Mascot search results for this experiment is shown in

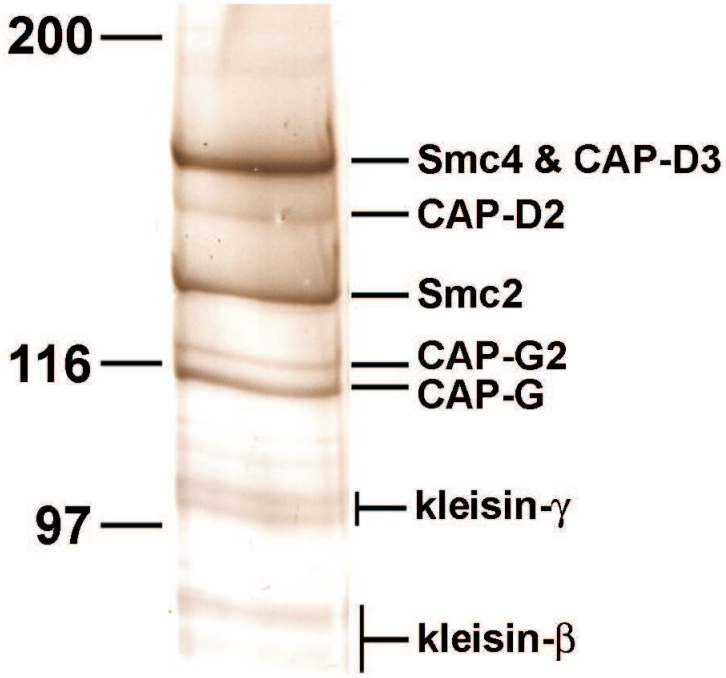


Figure 7.2: Quality of the condensin complex purification. SDSPAGE silver-stained gel of the purified human condensin complexes. The bands were previously identified by Yeong et al. [58]. This result confirms the purity of the complex obtained in the experiment.

Table 7.2.2. In the first the case of searching the small database consisting of 146 sequences was considered. Each of the eight condensin subunits showed an increase in Mascot score (mean increase of 8.2%), and number of peptide matches (mean increase of 4.8%) following the cleaning procedure.

As a rule, the percentage of sequence coverage obtained was the same or higher for searches using the cleaned spectra than for those using the original spectra. The only exception from this list was kleisin-β , which showed a 2% reduction in the sequence coverage after cleaning. Closer inspection revealed that this reduction was due to a single peptide match generated by a single MS/MS spectrum that visually appears of low quality (see data and figures at http://mendel.imp.ac.at/mass-spectrometry/falsepositive-partB.html). This MS/MS spectrum has very few significant peaks above the baseline, and is classified as "non-interpretable" by the MS Cleaner. We found out that this artifact is a result of low frequency suppression in the PSPS-graph and could be avoided with a slightly reduced threshold $f_{BT} = 12$. However, the Mascot program generated a match between this spectrum and the peptide QGEVLASR (within kleisin-β). It was classified as a hit with a Mascot score of 45, although the majority of the peaks that contributed to the assignment are very small and the most significant peaks do not contribute to this interpretation. Thus in this case, the removal of just a single non-reliable peptide during the cleaning process resulted in a small reduction in sequence coverage, although the Mascot score for the protein as a whole was increased as a result of background removal.

It should be noted that all cases of peptide detection by Mascot in spectra classified as "non-interpretable" by MS Cleaner (14 out 1318 dta-files) lead to low scores with marginal sequence coverage by Mascot when there are very few significant peaks above an apparent noise. Changing to Mascot

Protein	Raw			Cleaned			Increment			Bad		
	Score	Match	Cov(%)	Score	Match	Cov(%)	Score(%)	Match(%)	Cov(%)	Score(%)	Match(%)	Cov(%)
(A)												
Smc4	3768	329	57	4125	341	64	9.5	3.6	12.3	98	2	1
CAP-D2	3637	182	65	4038	195	69	11.0	7.1	6.2	33	1	1
Smc2	2957	219	55	3239	231	57	9.5	5.5	3.6	201	4	4
CAP-D3	2627	104	42	2772	108	43	5.5	3.8	2.4	n/a	n/a	n/a
CAP-G	2554	106	55	2678	110	55	4.9	3.8	0.0	200	3	3
CAP-G2	1992	82	44	2255	86	50	13.2	4.9	13.6	154	3	6
Kleisin-γ	1843	78	61	1979	84	63	7.4	7.7	3.3	n/a	n/a	n/a
Kleisin-β	1245	45	69	1306	46	67	4.9	2.2	-2.9	45	1	1
(B)												
Smc4	4829	416	62	5188	424	64	7.4	1.9	3.2			
CAP-D2	4411	229	66	4818	241	68	9.2	5.2	3.0			
Smc2	4054	300	61	4436	312	64	9.4	4.0	3.8			
CAP-D3	3134	118	43	3329	125	45	6.2	5.9	3.9			
CAP-G	2850	117	51	3014	120	52	5.8	2.6	1.5			
CAP-G2	2553	106	50	2760	110	51	8.1	3.8	1.8			
Kleisin-γ	2158	94	61	2300	96	61	6.6	2.1	0.7			
Kleisin-β	1446	48	65	1573	49	65	8.8	2.1	-0.8			
(C)												
Smc4	4502	321	59.860	4865	328	62	8.1	2.2	3.4			
CAP-D2	4176	192	64.954	4590	204	67	9.9	6.3	2.5			
Smc2	3747	246	59.733	4137	255	62	10.4	3.7	3.4			
CAP-D3	2862	100	53.695	3060	104	54	6.9	4.0	1.5			
CAP-G	2453	76	24.860	2627	81	25	7.1	6.6	2.5			
CAP-G2	2239	163	39.463	2500	165	41	11.7	1.2	3.4			
Kleisin-γ	1892	146	34.005	2167	149	36	14.5	2.1	5.9			
Kleisin-β	1043	31	45.785	1104	31	46	5.9	0.0	1.4			

Table 7.3: The MS/MS spectra were interpreted with MASCOT directly ("raw spectra" from 53 944 dta files, total size 460 MB) and after processing with the background removal procedure ("cleaned spectra" from 52 626 dta files, total size 284 MB) described in this article. The "score" is the MASCOT score from successful searches; "match" is the number of searches that recover the peptides from the protein used. "cov (%)" reports the sequence coverage. We present the results of three searches: (A) against the database of 146 proteins, (B) against the human proteins in the nonredundant database and (C) against all proteins in the nonredundant database. The columns "bad spectra" report cases of files (among 1318 dta files, total size 7 MB) that are considered not interpretable by the criterion described in the text (n/a - not applicable) where MASCOT could, nevertheless, recognize the original protein in a database of 146 proteins but with a low score. Cov., Coverage.

searches against larger databases leads, as a trend, to even more dramatic improvements of scores and sequence matches (Table 7.2.2). In the case of the full non-redundant protein sequence database, there is even an increase of sequence coverage for kleisin-β after background removal with our procedure because Mascot was unable to assign a match to several noisy spectra against the extensive sequence background of the largest database.

In a practical setup, the computational efficiency is also important. MS Cleaner processed the 53944 spectra from the condensin experiment in less than 4 hours on a single standard PC; i.e., in 0.25 seconds per file. However, the application of our background removal procedure reduces the pure Mascot computing time for the body of 53944 dta-files in the condensin complex case by about 25%, even in the case of a small database of 146 sequences; the size of the cleaned mgf-file is decreased by 39%. Therefore, application of the MS Cleaner significantly reduces consumption of computing time and storage.

7.2.3 Comparison Between Mascot Distiller and MS Cleaner

There are no tools for background removal in peptide MS/MS spectra readily available in the public domain. Among commercial programs, only Mascot Distiller is explicitly devoted to this task. From the scientific point of view, a correct comparison of Mascot Distiller with our tool is not possible, because the algorithms used in commercial Mascot Distiller have not been properly described in public and the reasons for differential performance of the two programs cannot be causally interpreted. Table 7.2.3 shows the results of application of the two programs on the BSA-, ADH- and TRF-datasets.

Whereas Mascot Distiller produces mixed results with respect to the score and sequence matches (one increase, two decreases), our program increases

Protein	Raw		Mascot Distiller			MS Cleaner		
	Score	Match	Score	Match	Time	Score	Match	Time
BSA	1844	65	1565	44	7:40	2094	70	3:58
ADH	536	24	612	15	6:48	594	25	2:34
TRF	1643	61	1532	38	5:48	1846	65	3:23

Table 7.4: The MS/MS spectra for BSA, ADH, and TRF were interpreted with MASCOT directly ("raw spectra") and after processing with MASCOT Distiller and with the background removal procedure described in this article ("MS Cleaner"). The "score" is the MASCOT score from all successful searches; "match" is the number of searches that recover the peptides from the protein used. The processing time is presented in min:sec. The performance of the procedure described in this article is superior compared with that of MASCOT Distiller with respect to score, and number of correct sequence matches. In addition, it consumes only 50% time on an identical computer with the same operating system environment.

the score and the number of matches in all three cases. At the same time, the computation time is only about 50% of that from Mascot Distiller. In the case of the larger condensin dataset, Mascot Distiller did not complete computation regularly and interrupted with a run-time error. As was shown above, application of our software improved the interpretability of the condensin dataset.

7.3 Tests on the Detection of Large Number of Non-Interpretable Spectra Using Sequence Ladder Length and Peak Intensity Threshold

Detection of non-interpretable spectra within MS Cleaner is carried out by two independent procedures. The Fourier-transform-based algorithm described in section 5.11.1 recognizes only a small number of bad spectra (below 1% of the total raw spectra). The sequence ladder test (see section 5.2) is highly efficient in removing non-interpretable spectra as the results described below convincingly show. For its practical application, it is necessary to determine two parameters. For the estimation of their optimal values, a systematic analysis on more than 270 000 of spectra was performed. Sequence ladder length was tested with values between 2 and 6; and intensity threshold ranges from 5% to 35%.

The results of a parameters subset are presented in Table 7.3 and Table 7.3. According to the expectations, the number of detected bad spectra increased with increasing sequence ladder length and decreasing intensity threshold (TableTable 7.3). The removal of bad spectra by the sequence ladder test decreases Mascot computation time with almost unchanged sequence coverage. Mascot scores increase due to the significance of the interpretation result obtained from a smaller set of peaks within the spectra.

To detect most of the bad spectra and save the interpretation time, the parameters are suggested as shown in the Table 7.3. With sequence ladder length equal 4 and intensity threshold of 20%, it is possible to eliminate up to ≈90% of all spectra (in average ≈65%) by declaring them as non-

interpretable spectra. The minor sequence coverage loss observed in only a few cases (BSA and ADH in TableTable 7.3) doesn't affect the interpretation result.

In the cases of small datasets (BSA, ADH and TRF), it was possible to run Mascot on a single-processor PC as standalone application and to measure the total computation time for interpretation (Table 7.3). The data shows that the interpretation time narrows up to only ≈20% of the original computation time if the intensity threshold 20% is applied. For the remaining larger datasets, computation was only possible on a larger Linux cluster in parallel calculation with other jobs; thus, the exact determination of the computation time required was not possible. Since the reduction of computation time required by Mascot is roughly proportional to the number of MS/MS spectra to be interpreted and the size of the dataset in bytes, we think that the savings of computation time for the other datasets are in the same order of magnitude.

It can been seen in Table 7.3 and Table 7.3 that the number of spectra classified as non-interpretable depends on severity of the parameters "sequence ladder length" and "intensity threshold". Nevertheless, even more relaxed parameters settings compared with the parameter pair (4; 20%) show considerable background removal capability. Therefore, if the sequence coverage is more important than computational time savings, softer parameters can be chosen with intensity threshold of 25%.

The columns A1-A16 from Table 7.3 have the following meaning:

A1 Name of the mass spectrometric dataset,

A2 Number of MS/MS spectra,

A3 Mascot Score obtained before background removal with MS Cleaner,

Protein	Sequence ladder length	Intensity threshold [%]	Cleaned spectra	Bad spectra	MS Cleaner time [min]	Mascot time [min]	Mascot score	Queries matched	Sequence coverage
BSA	0	100	-	-	-	61	586	89	55
	3	100	1664	1015	3.92	44	720	91	57
	3	15	390	2289	1.21	17	1991	84	52
	3	20	490	2189	1.40	21	2108	87	57
	3	25	601	2078	1.61	26	2114	89	57
	3	30	688	1991	1.75	29	2114	90	57
	4	100	940	1739	3.80	36	2108	91	57
	4	15	260	2419	0.91	12	1875	78	47
	4	20	321	2358	1.06	14	1911	80	47
	4	25	380	2299	1.25	18	2114	86	57
	4	30	441	2238	1.30	19	2114	89	57
	5	100	593	2086	3.82	26	2108	91	57
	5	15	174	2505	0.60	9	1579	60	41
	5	20	232	2447	0.85	11	1809	72	44
	5	25	281	2398	1.00	13	1963	81	49
	5	30	313	2366	0.85	14	2058	86	54
ADH	0	100	-	-	-	64	242	39	39
	3	100	1446	879	4.15	45	327	34	39
	3	15	269	2056	0.88	12	673	29	35
	3	20	347	1978	1.10	13	696	31	37
	3	25	440	1885	1.33	17	697	32	37
	3	30	697	1628	1.53	20	697	33	37
	4	100	902	1423	4.15	35	733	34	39
	4	15	173	2152	0.58	7	562	26	28
	4	20	216	2109	0.71	9	673	30	35
	4	25	271	2054	0.90	12	607	28	33
	4	30	325	2000	1.05	13	697	32	37
	5	100	594	1731	4.20	23	712	33	39
	5	15	94	2231	0.35	5	311	15	21
	5	20	125	2200	0.46	6	366	17	25
	5	25	145	2180	0.53	7	434	19	26
	5	30	186	2139	0.66	9	589	24	31
TRF	0	100	-	-	-	52	588	86	47
	3	100	1587	1021	3.57	42	768	87	49
	3	15	373	2235	1.00	17	1988	86	49
	3	20	485	2123	1.23	20	1988	86	49
	3	25	568	2040	1.36	24	1998	87	49
	3	30	639	1969	0.78	27	1998	87	49
	4	100	864	1744	3.62	34	1973	87	49
	4	15	231	2377	0.70	11	1987	81	49
	4	20	298	2310	0.86	13	1988	84	49
	4	25	360	2248	1.00	16	1988	85	49
	4	30	414	2194	1.12	19	1998	87	49
	5	100	540	2068	3.63	23	1973	87	49
	5	15	164	2444	0.55	9	1785	68	45
	5	20	194	2414	0.61	10	1890	74	47
	5	25	245	2363	0.75	12	1957	80	48
	5	30	286	2322	0.86	14	1968	84	48

Table 7.5: Influence of background removal on the recovery of MS/MS spectra of 100 fmol test samples. The original number of MS/MS spectra for the BSA, ADH and TRF datasets are 2679, 2325 and 2608 respectively. The intensity threshold (column 3) describes the search of the sequence ladder (column 2) within the 15%, 20%, 25% or 30% top peaks (100% - all peaks are considered). The following three columns show the MS Cleaner output - number of spectra with background removal, of unselected spectra and the MS Cleaner CPU time on a single-processor Windows XP computer (Pentium IV 2.4 GHz, 1G RAM). The remaining four columns present the MASCOT output - the CPU time on the same machine, the protein score, the number of matching spectra and the final sequence coverage. For each dataset, the first line shows the results for the case when MS Cleaner is not used for pre-processing and the MS/MS data is immediately interpreted by MASCOT.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
BSA	2679	586	55	64.91	2108	57	3.80	88.02	1911	47	1.06	85.82	2114	57	1.25
ADH	2325	242	39	61.20	733	39	4.15	90.71	673	35	0.71	88.34	607	33	0.90
Transferin	2608	588	47	66.87	1973	49	3.61	88.57	1988	49	0.86	86.20	1988	49	1.00
AlphaAmyl-col1	10108	633	24	11.30	667	24	31.65	60.07	667	24	15.13	51.09	667	24	18.07
AlphaAmyl-col2	10184	698	35	9.82	780	35	34.20	50.22	780	35	19.05	20.25	780	35	22.76
AmylGlu-col1	10030	736	28	13.26	761	28	28.40	79.24	761	28	8.66	73.58	761	28	10.63
AmylGlu-col2	9870	801	36	13.31	860	37	29.50	72.62	860	37	11.70	63.95	860	37	14.29
Apo-col1	10032	2606	63	11.72	2814	63	30.76	63.10	2814	63	13.93	54.49	2814	63	16.78
Apo-col2	10090	2571	60	12.13	2761	60	32.95	53.12	2761	60	17.53	44.32	2761	60	21.03
BetaGal-col1	10324	1459	56	7.17	1567	57	34.98	48.06	1567	57	22.05	40.53	1567	57	24.60
BetaGal-col2	10368	1309	51	8.12	1508	56	36.71	42.90	1454	55	24.76	33.10	1454	55	28.61
CarAnly-col1	9946	586	49	12.35	616	49	26.35	90.31	573	49	3.65	84.94	607	49	5.48
CarAnly-col2	9534	582	52	13.40	616	52	26.27	86.07	616	52	5.08	78.44	616	52	7.66
Cat-col1	10098	1798	61	11.13	1886	61	30.88	67.26	1879	61	13.13	57.89	1879	61	16.50
Cat-col2	10034	1567	65	11.78	1693	65	31.90	59.50	1693	65	15.91	48.55	1693	65	19.56
PhosB-col1	10118	2780	59	10.30	3079	61	35.13	63.49	3014	60	14.26	54.46	3047	61	17.25
PhosB-col2	10096	2655	61	10.52	3116	65	32.58	53.96	3084	65	17.58	44.31	3116	65	21.16
GluDey-col1	10006	892	36	11.29	986	36	27.30	79.55	986	36	7.75	73.42	986	36	9.71
GluDey-col2	9886	850	34	11.81	962	34	28.73	72.51	962	34	10.13	62.25	962	34	13.51
GluTra-col1	10022	351	25	10.36	389	25	28.61	71.64	348	25	10.25	62.78	389	25	14.30
GluTra-col2	10156	341	33	9.18	384	33	31.31	61.15	384	33	14.25	49.59	384	33	28.11
Immo-col1	10330	506	35	9.27	565	35	36.20	42.30	565	35	24.95	34.44	565	35	27.66
Immo-col2	10334	356	66	8.61	500	66	38.05	37.06	500	66	27.31	28.47	500	66	30.31
LacDe-col1	10286	1549	58	10.36	1694	58	35.36	53.20	1694	58	20.03	44.86	1694	58	23.15
LacDe-col2	10250	1346	54	9.07	1483	54	36.48	40.16	1483	54	25.60	31.67	1483	54	28.31
LactoPee-col1	10242	1613	45	13.16	1764	45	34.78	62.12	1756	45	15.91	52.37	1764	45	19.53
LactoPee-col2	10402	1679	43	9.09	1890	44	35.18	51.70	1890	44	20.31	41.76	1890	44	23.85
Myo-col1	9958	561	66	11.67	594	66	27.26	85.42	594	66	5.46	79.25	594	66	7.45
Myo-col2	9744	530	66	12.15	584	66	28.01	80.83	584	66	6.95	70.92	584	66	10.35

Table 7.6: Large scale testing of the sequence ladder test as implemented in MS Cleaner.

A4 Sequence Coverage before background removal with MS Cleaner [% of original target sequence length],

A5 Non-interpretable spectra detected when applying the sequence ladder length 4 and the intensity threshold 100% [% of the number of MS/MS spectra in column A2],

A6 Mascot Score found when applying the sequence ladder length 4 and the intensity threshold 100%,

A7 Sequence Coverage found when applying the sequence ladder length 4 and the intensity threshold 100% [% of original target sequence length],

A8 MS Cleaner computation time [min] when applying the sequence ladder length 4 and the intensity threshold 100%,

A9 Non-interpretable spectra detected when applying the sequence

ladder length 4 and the intensity threshold 20% [% of the number of MS/MS spectra in column A2],

A10 Mascot Score found when applying the sequence ladder length 4 and the intensity threshold 20%,

A11 Sequence Coverage found when applying the sequence ladder length 4 and the intensity threshold 20% [% of original target sequence length],

A12 MS Cleaner computation time [min] when applying the sequence ladder length 4 and the intensity threshold 20%,

A13 Non-interpretable spectra detected when applying the sequence ladder length 4 and the intensity threshold 25% [% of the number of MS/MS spectra in column A2],

A14 Mascot Score found when applying the sequence ladder length 4 and the intensity threshold 25%,

A15 Sequence Coverage found when applying the sequence ladder length 4 and the intensity threshold 25% [% of original target sequence length],

A16 MS Cleaner computation time [min] when applying the sequence ladder length 4 and the intensity threshold 25%.

MS/MS dataset	Raw data	Intensity threshold 100%	Intensity threshold 20%	Intensity threshold 25%
BSA	61	36	14	18
ADH	64	35	9	12
TRF	52	42	20	24

Table 7.7: Computation time required for the interpretation of MS/MS datasets with Mascot. In this table, the computation time consumed by Mascot (min) is contrasted for the case of interpreting the untreated data (column 2) with the cases of application of the sequence ladder test with sequence ladder length 4 and varying intensity thresholds.

Chapter 8

Conclusions

In this work, it was shown that it is possible to recognize a considerable amount of background noise in tandem mass spectra of peptides. First of all, spectra that are non-interpretable as amino acid sequences can be filtered out with the sequence ladder test. Large scale testing over mass spectrometry datasets of proteins show that this criterion effectively removes about 65% of the spectra as non-relevant for protein identification. Spectra that contain important peptide information still comprise large quantities of noise peaks. Cases of multiply charged isotope clusters can be recognized with an etalon-correlation method if the data is accurately recorded. Then, the isotope cluster can be removed and substituted by correct monoisotopic peak with single charge level. Spectral analysis methods known from the signal processing theory can effectively be used to eliminate oddities in the frequency spectrum of the MS/MS spectrum (latent periodicities and high-frequency components) and, in this way, remove a considerable number of non-relevant peaks. In average, spectra are reduced by one quarter in size.

This processing of peptide MS/MS spectra positively affects protein identification. Not only does the procedure essentially not lead to any loss of

information, interpretation success rate and reliability is improved in many cases.

The program MS Cleaner, the implementation of the algorithms described in this work, is suggested to be used as routine pre-processing procedure in mass spectrometric applications in the proteomics field.

The results of this work have partially been published in an article of the journal "Protemics" [33].

Bibliography

[1] Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Boucherie H, Mann M., *Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels*, Proc Natl Acad Sci USA 1996;93:14440-14445.

[2] Pandey A, Mann M., *Proteomics to study genes and genomes*, Nature 2000;405:837-846.

[3] McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates JR., *Direct Analysis and Identification of Proteins in Mixtures by LC/MS/MS and Database Searching at the Low-Femtomole Level*, Anal Chem 1997;69:767-776.

[4] Washburn MP, Wolters D, Yates JR., *Large-scale analysis of the yeast proteome by multidimensional protein identification technology*, Nat Biotechnol 2001;19:242-247.

[5] Wysocki VH, Tsaprailis G, Smith LL, Breci LA., *Mobile and localized protons: a framework for understanding peptide dissociation*, J Mass Spectrom 2000;35:1399-1406.

[6] Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR., *Protein Sequencing by Tandem Mass Spectrometry*, Proceedings of the National Academy of Sciences 1986;83:6233-6237.

[7] Poulter L, Tylor LC. Int J Mass Spectrom Ion Processes 1989;91:183-197.

[8] Alexander AJ, Thibault P, Boyd RK, Curtis JM, Rinehart KL. Int J Mass Spectrom Ion Processes 1990;98:107-134.

[9] Somogyi A, Wysocki VH, Mayer I. J Am Soc Mass Spectrom 1994;5:704-717.

[10] Papayannopoulos IA. Mass Spectrom Rev 1995;14:49-73.

[11] Cox KA, Gaskell SJ, Morris M, Whiting A. J Am Soc Mass Spectrom 1996;7:522-531.

[12] Dongre AR, Jones JL, Somogyi A, Wysocki VH. J Am Soc Mass Spectrom 1996;118:8365-8374.

[13] Yergey J, Heller D, Hansen G, Cotter RJ, Fenselau C., *Isotopic Distributions in Mass Spectra of Large Molecules*, Anal Chem 1983;55:353-356.

[14] Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Science 1989;246:64-71.

[15] Mann M. Org Mass Spectrom 1990;25:575-587.

[16] Smith RD, Loo JA, Ogorzalek Loo RR, busman M, Udseth HR. Mass Spectrom Rev 1991;10:359-451.

[17] Kebarle P, Tang L. Anal Chem 1993;65:972A-986A.

[18] McLafferty FW. Acc Chem Res 1994;27:379-386.

[19] Scoble HA, Biller JE, Biemann K., *A graphics display-oriented strategy for the amino acid sequencing of peptides by tandem mass spectrometry*, Fresenius Z Anal Chem 1987;327:239-245.

[20] Bartels C., *Fast Algorithm for Peptide Sequencing by Mass Spectroscopy*, Biomed Environ Mass Spectrom 1990;19:363-368.

[21] Johnson RS, Taylor JA., *Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry*, Mol Biotechnol 2002;22:301-315.

[22] Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA, *De novo peptide sequencing via tandem mass spectrometry*, J Comput Biol 1999;6:327-342.

[23] Zhang Z, McElvain JS., *De Novo Peptide Sequencing by Two-Dimensional Fragment Correlation Mass Spectrometry*, Anal Chem 2000;72:2337-2350.

[24] Horn DM, Zubarev RA, McLafferty FW., *Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry*, PNAS 1994;97:10313-10317.

[25] Taylor JA, Johnson RS., *Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry*, Anal Chem 2001;73:2594-2604.

[26] Eng JK, McCormack AL, Yates JR., *An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database*, J Am Soc Mass Spectrom 1994;5:976-989.

[27] Yates JR, Eng J, McCormack AL, Schieltz DM., *Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database*, Anal Chem 1995;67:1426-1436.

[28] Yates JR, III, McCormack AL, Eng J., *Mining genomes with MS*, Anal Chem 1996;68:534A-540A.

[29] Yates JR, III, Eng JK, McCormack AL., *Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases*, Anal Chem 1995;67:3202-3210.

[30] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS., *Probability-based protein identification by searching sequence databases using mass spectrometry data*, Electrophoresis 1999;20:3551-3567.

[31] Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ, Yates JR, III., *Code developments to improve the efficiency of automated MS/MS spectra interpretation.*, J Proteome Res 2002;1:211-215.

[32] Zhang N, Aebersold R, Schwikowski B., *ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data*, Proteomics 2002;2:1406-1412.

[33] Mujezinovic N, Raidl G, Hutchins JR, Peters JM, Mechtler K, Eisenhaber F., *Cleaning of raw peptide MS/MS spectra: improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise*, Proteomics 2006;6:5117-5131.

[34] Mann M, Meng CK, Fenn JB., *Interpreting mass spectra of multiply charged ions*, Anal Chem 1989;61:1702-1708.

[35] Ferrige AG, Seddon MJ., *Maximum Entropy Deconvolution in Electrospray Mass Spectrometry*, Rapid Commun Mass Spectrom 1991;5:374-379.

[36] Reinhold BB, Reinhold VN., *Electrospray ionization mass spectrometry: Deconvolution by an entropy-based algorithm*, J Am Soc Mass Spectrom 1992;3:207-215.

[37] Zhang Z, Marshall A., *A Universal Algorithm for Fast and Automated Charge State Deconvolution of Electrospray Mass-to-Charge Ratio Spectra*, J Am Soc Mass Spectrom 1998;9:225-233.

[38] Gentzel M, Kocher T, Ponnusamy S, Wilm M., *Preprocessing of tandem mass spectrometric data to support automatic protein identification*, Proteomics 2003;3:1597-1610.

[39] Wehofsky M., *Struktureinflsse auf das Fragmentierungs-Verhalten von Peptiden bei PSD-MALDI Massenspektrometrie*, In: Justus-Liebig-Universitt Giessen, Germany; 2001.

[40] Wehofsky M, Hoffmann R., *Automated deconvolution and deisotoping of electrospray mass spectra*, J Mass Spectrom 2002;37:223-229.

[41] Jaitly D, Page-Belanger R, Faubert D, Thibault P, Kebarle P., *MSMS Peak Identification and its Applications* In: 2004. 3 p.

[42] Horn DM, Zubarev RA, McLafferty FW., *Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules*, J Am Soc Mass Spectrom 2000;11:320-332.

[43] Xu M, Geer LY, Bryant SH, Roth JS, Kowalak JA, Maynard DM, Markey SP., *Assessing data quality of Peptide mass spectra obtained by quadrupole ion trap mass spectrometry*, J Proteome Res 2005;4:300-305.

[44] Bern M, Goldberg D, McDonald WH, Yates JR, III., *Automatic quality assessment of Peptide tandem mass spectra*, Bioinformatics 2004;20 Suppl 1:I49-I54.

[45] F.Lottspeich, H.Zorbas. Bioanalytik. 1998.

[46] Yamashita M, Fenn JB., *Electrospray ion source. Another variation on the free-jet theme*, J Phys Chem 1984;88:4451-4459.

[47] Wilm M, Mann M., *Analytical Properties of the Nanoelectrospray Ion Source*, Anal Chem 1996;68:1-8.

[48] Hillenkamp F, Karas M, Beavis RC, Chait BT., *Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers*, Anal Chem 1991;63:1193A-1203A.

[49] Roepstorff P, Fohlman J., *Proposal for a common nomenclature for sequence ions in mass spectra of peptides*, Biomed Mass Spectrom 1984;11:601.

[50] Johnson RS, Biemann K., *Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides*, Biomed Environ Mass Spectrom 1989;18:945-957.

[51] Ashcroft AE, Derrick PJ., *Four-Sector Tandem Mass Spectrometry of Peptides. In. Mass Spectrometry of Peptides*, Florida: CRC Press; 1990.

[52] Hirota T, Gerlich D, Koch B, Ellenberg J, Peters JM., *Distinct functions of condensin I and II in mitotic chromosome assembly*, J Cell Sci 2004;117:6435-6445.

[53] Mitulovic G, Smoluch M, Chervet JP, Steinmacher I, Kungl A, Mechtler K., *An improved method for tracking and reducing the void volume in nano HPLC-MS with micro trapping columns*, Anal Bioanal Chem 2003;376:946-951.

[54] Mitulovic G, Smoluch M, Chervet JP, Steinmacher I, Kungl A, Mechtler K., *An improved method for tracking and reducing the void volume in nano HPLC-MS with micro trapping columns*, Anal Bioanal Chem 2003;376:946-951.

[55] Mitulovic G, Stingl C, Smoluch M, Swart R, Chervet JP, Steinmacher I, Gerner C, Mechtler K., *Automated, on-line two-dimensional nano liquid chromatography tandem mass spectrometry for rapid analysis of complex protein digests*, Proteomics 2004;4:2545-2557.

[56] Oppenheim AV, R.W.Schafer. *Discrete-Time Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall; 1989.

[57] Blom KF., *Elemental composition from moment analysis of the low-resolution isotope pattern*, Org Mass Spectrom 1988;23:194-203.

[58] She J, McKinney M, Petreas M, Stephens R., *Design and application of an isotope pattern calculator for Microsoft Windows*, Organohalogen Compd 1995;23:171-174.

[59] Rockwood AL., *Relationship of Fourier-Transforms to Isotope Distribution Calculations*, Rapid Commun Mass Spectrom 1995;9:103-105.

[60] Rockwood AL, VanOrden SL., *Ultrahigh-speed calculation of isotope distributions* Anal Chem 1996;68:2027-2030.

[61] Rockwood AL, VanOrden SL, Smith RD., *Ultrahigh resolution isotope distribution calculations*, Rapid Commun Mass Spectrom 1996;10:54-59.

[62] Wehofsky M, Hoffmann R., *Automated deconvolution and deisotoping of electrospray mass spectra*, J Mass Spectrom 2002;37:223-229.

[63] Wehofsky M, Hoffmann R., *Automated deconvolution and deisotoping of electrospray mass spectra*, J Mass Spectrom 2002;37:223-229.

[64] Baranov V., *Method for reducing chemical background in mass spectra*, In. USA/CA; 2003.

[65] Friedlander B, Porat B., *The Modified Yule-Walker Method of ARMA Spectral Estimation*, IEEE Transactions on Aerospace Electronic Systems 1984;AES-20:158-173.

[66] Parks TW, Burrus CS., *Digital Filter Design*, In. New York: John Wiley and Sons; 1987.

[67] Oppenheim AVaRWS., *Discrete-Time Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall; 1989. 312 p.

[68] Hirota T, Gerlich D, Koch B, Ellenberg J, Peters JM., *Distinct functions of condensin I and II in mitotic chromosome assembly*, J Cell Sci 2004;117:6435-6445.

[69] Ono T, Losada A, Hirano M, Myers MP, Neuwald AF, Hirano T., *Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells*, Cell 2003;115:109-121.

[70] Schleiffer A, Kaitna S, Maurer-Stroh S, Glotzer M, Nasmyth K, Eisenhaber F., *Kleisins: a superfamily of bacterial and eukaryotic SMC protein partners*, Mol Cell 2003;11:571-575.

[71] Yeong FM, Hombauer H, Wendt KS, Hirota T, Mudrak I, Mechtler K, Loregger T, Marchler-Bauer A, Tanaka K, Peters JM, Ogris E., *Identification of a subunit of a novel Kleisin-beta/SMC complex as a potential substrate of protein phosphatase 2A*, Curr Biol 2003;13:2058-2064.

[72] N. Brenner and C. Rader, 1976, *A New Principle for Fast Fourier Transformation*, IEEE Acoustics, Speech and Signal Processing 24: 264-266.

[73] Cooley, James W., and John W. Tukey, 1965, *An algorithm for the machine calculation of complex Fourier series*, Math. Comput. 19: 297301.

[74] MATLAB, *Signal Processing Toolbox*, MathWorks Inc.

Curriculum Vitae

Personal Information

Name: Nedim Mujezinovic
Date and place of birth: November 24^{th} 1970 in Tuzla, Bosnia and Herzegovina
Nationality: Bosnia and Herzegovina

Education

1977 – 1985	Primary School "Vojo Ivanovic Crnogorac", Tuzla, Bosnia and Herzegovina
1985 – 1989	Secondary School "Mesa Selimovic", Tuzla, Bosnia and Herzegovina ("Class for Informatics, Mathematics and Physics", Profession obtained: Programmer)
1997 – 2003	M.Sc. of Biochemistry and Biochemical Engineering at the Vienna University of Technology
2003 – 2004	M.Sc. Thesis: "Deficiencies of current protein MS/MS spectra evaluation techniques and strategies for their methodical improvement", Bioinformatics Group, IMP//Vienna
2003	Microsoft Certified Professional, "Developing and imple-

	menting Desktop Applications with Visual C++"
2004 – 2007	Ph.D. Student, Thesis title: "Improved Protein Identification after Fast Elimination of Non-Interpretable Peptide MS/MS Spectra and Noise Reduction" at the Bioinformatics Group and Protein Chemistry Facility, IMP Vienna, and Institute for Algorithms and Data Structures, Vienna University of Technology.

Working experience

1992 – 1995	IT Specialist, Bosnian Army
1996 – 1997	IT Manager and Database Administrator, OSCE Mission to Bosnia and Herzegovina
2002 – 2003	Senior Software Developer, A-Null GMbH, Vienna
2004 – 2005	Lecturer, FH Campus Wien (Bioinformatics)

Scientific Publications

2005	US Patent, title: "Method for removal of multiply charged peaks, isotope replicates, periodic and high-frequency noise from protein MS/MS spectra as well as method for the recognition of non-interpretable protein MS/MS spectra"
2006	Mujezinovic, N. et al. "Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise." Proteomics. 6, 5117-5131 (2006).
In preparation	Mujezinovic, N. et al. "Reducing the haystack to find the

needle: Improved protein identification after fast elimination of non-interpretable peptide MS/MS spectra and noise reduction".

Languages Bosnian (mother tongue), English, German