EXPLOITING FRAGMENT-ION COMPLEMENTARITY FOR PEPTIDE *DE NOVO* SEQUENCING FROM COLLISION INDUCED DISSOCIATION TANDEM MASS SPECTRA

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Molecular Biology and Genetics

by

Belgin AYTUN

July 2011
İZMİR

We approve the thesis of **Belgin AYTUN**

Assist. Prof. Dr. Jens ALLMER
Supervisor

Assoc. Prof. Dr. Talat YALÇIN
Committee Member

Assist. Prof. Dr. Ferda SOYER
Committee Member

1 July 2011

Assoc. Prof. Dr. Ahmet KOÇ
Head of the Department of Molecular
Biology and Genetics

Prof. Dr. Durmuş Ali DEMİR
Dean of the Graduate School of
Engineering and Science

ACKNOWLEDGMENTS

I am highly indebted to my supervisor Assist. Prof. Dr. Jens ALLMER. He suggested the topic of this thesis and provided valuable feedback and scientific input for this work. He also taught me a lot about mass spectrometry and bioinformatics in general. His invaluable motivation, advice and comments have the largest immediate influence on the thesis. I am deeply grateful for ALLMER's valuable scientific advice I could always rely on. I am very happy to have been able to work with him.

Also I would like to express my grateful thanks to my committee members Assoc. Prof. Dr. Talat YALÇIN and Assist. Prof. Dr. Ferda SOYER for their suggestions.

I would like to thank Assist. Prof. Dr. Tuğkan TUĞLULAR. Especially in the early months, he spent endless hours teaching me how to program JAVA.

I am also grateful to my colleagues, Canan HAS, Şule YILMAZ and Visam GÜLTEKİN for their sincere motivation, help and kindness during my studies.

Last but not by any means least, I am more indebted to my father, mother, sister and fiance than I am able to express in words. Their endless love and unwavering support for my pursuit of ambitious and faraway goals continues to be a fundamental pillar in my ability to undertake and achieve these goals. You may be sometimes far but you are never absent.

ABSTRACT

EXPLOITING FRAGMENT-ION COMPLEMENTARITY FOR PEPTIDE *DE NOVO* SEQUENCING FROM COLLISION INDUCED DISSOCIATION TANDEM MASS SPECTRA

Peptide identification from mass spectrometric data is a key step in proteomics because this field provides sequence, quantitative, and modification data of actually expressed proteins. Two approaches are generally deployed to interpret experimental MS/MS data, database searching and *de novo* sequencing. Database search method has been used successfully in proteomics projects for organisms with well-studied genomes. However, it is not applicable in situations where a target sequence is not in the protein database. This can happen for a number of reasons, including novel proteins, protein mutations and post-translational modifications. Because of the disadvantages of database searching method, a lot of research has focused on *de novo* sequencing method which assigns amino acid sequences to MS/MS spectra without the need for a database.

The aim of this study is to enhance the accuracy of *de novo* sequencing tools. One step commonly employed in all *de novo* sequencing tools is naming of fragment ions. It is essential to know which peak represents which ion type in order to traverse a spectrum graph to find an amino acid sequence that best explains the MS/MS spectrum. Different approaches have been tried to name ions and some success has been achieved in naming b-type ions and y-type ions.

We have presented a new approach which enables the naming of not only b- and y-type ions but other arbitrary ion types as well. This enabled the detection of b-ion ladder. In the latter case, missing fragments were determined by using other named ion types. Furthermore, unexplained data in tandem mass spectra were reduced as much as possible. Therefore, a complete sequence will be derived by the new approach.

ÖZET

PEPTİT *DE NOVO* DİZİLİMİ İÇİN ÇARPIŞMALI-İNDÜKLENMİŞ TANDEM KÜTLE SPEKTRUMLARINDAKİ PARÇALANMIŞ İYONLARIN TAMAMLAYICILIĞINDAN YARARLANILMASI

Kütle spekrumu verilerinden peptit tanımlaması proteomik çalışmalarında temel bir adımdır. Çünkü bu alan, ifade olmuş proteinin dizi, niceliksel ve modifikasyon çalışmalarında kullanıldığından biyolojik sistemlerin anlaşılmasına katkı sağlar. İki yaklaşım deneysel tandem kütle spektrum verilerine uygulanır, bunlar veritabanı araması ve *de novo* dizilemedir. Veritabanı arama metodu, genomu tümüyle çalışılmış organizmaların proteomik projelerinde etkili bir şekilde kullanılmaktadır. Fakat, aranan peptit dizisi protein veritabanında bulunmuyor ise bu yaklaşım uygulanamaz. Veritabanında bulunmama bazı nedenlerden dolayı olabilir. Örneğin, çalışılan protein yeni ise, mutasyon ve tranlasyon sonrası modifikasyon içeriyorsa protein veritabanında bulunmaz. Bu dezavantaj nedeniyle, birçok araştırma *de novo* dizilimi metoduna odaklanmıştır. Çünkü, *de novo* dizilimi metodu tandem kütle spektrumlarınn peptide dizisini hiçbir veritabanına vereksinim duymadan belirleyebilir.

Bu çalışmanın amacı, *de novo* dizileme algoritmalarının doğruluğunun geliştirilmesidir. Fragment iyonlarını isimlendirme tüm *de novo* dizileme algoritmalarnın (genellikle ima edilir) ortak bir adımıdır. Çünkü, spektrum grafik kullanarak tandem kütle spektrumunu en iyi ifade eden amino asit dizisini bulmak için, fragment iyonlarının hangi iyon tipini gösterdiğinin bilinmesi gerekmektedir. Farklı yaklaşımlar iyon isimlendirme için denenmiştir ve bazıları b-tipi iyonları ve y-tipi iyonlarını isimlendirerek başarılı olmuştur.

Yeni bir yaklaşımla, sadece b-iyon tipi yada y-iyon tipi isimlendirerek değil, diğer rastgele seçilmiş iyon tiplerini de isimlendirerek sonuçta b-iyon tipi merdivenini saptamayı hedefledik. Ayrıca, eksik ve peptit dizilemede önemli olan iyonları da diğer isimlendirilmiş iyon tipleriyle bazı spektrumlarda saptadık. Bunlara ek olarak, tandem kütle spektralarındaki anlaşılmayan veriler mümkün olduğunca azaltılmıştır. Böylece, b-iyon tipi merdivenini kullanarak tam peptit dizisini bulmak mümkün olabilecekiır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. Proteins, Proteome, and Proteomics

In the 19th century, proteins were discovered and named by Jöns Jakob Berzelius. Proteins are constructed from a set of 20 amino acids which are linked to each other by peptide bond. For all living organisms proteins have a crucial role. They can act as enzymes which accelerate many important chemical reactions. Proteins also form the backbone of the cells and different cellular structures and participate in vital roles such as mitosis, cell motility and signaling. Their involvement in most of the cellular processes and functions highlights the significance of the proteins (Filén, 2008).

The protein synthesis in the cells is expressed according to the genetic code. Except in some viruses as RNA, genetic information is stored in DNA. A gene which represents a DNA sequence codes the corresponding protein. In protein synthesis, at first the DNA is transcribed to a mRNA transcript which is, then, further translated into a protein.

In molecular biology, the central dogma refers to the relationship between DNA, mRNA and protein (Varki et al., 1999). The central dogma has been expanded due to existing of post-translational modification and alternative splicing. Post-translational modification (PTM) is the chemical modification of a protein after its translation. For various proteins, PTM is one of the later step in protein biosynthesis (Santos-Rosa et al., 2002). Alternative splicing is a process by which the exons of the RNA produced by transcription of a gene are reconnected in multiple ways during RNA splicing (Leipzig, Pevzner, & Hever, 2004). The resulting different mRNAs may be translated into different protein isoforms. Therefore a single gene may code for multiple proteins (Pandey & Mann, 2000). The relationship between DNA, mRNA, alternative splicing, protein, posttranslational modification is referred as an expended central dogma of biology, see Figure 1.1.

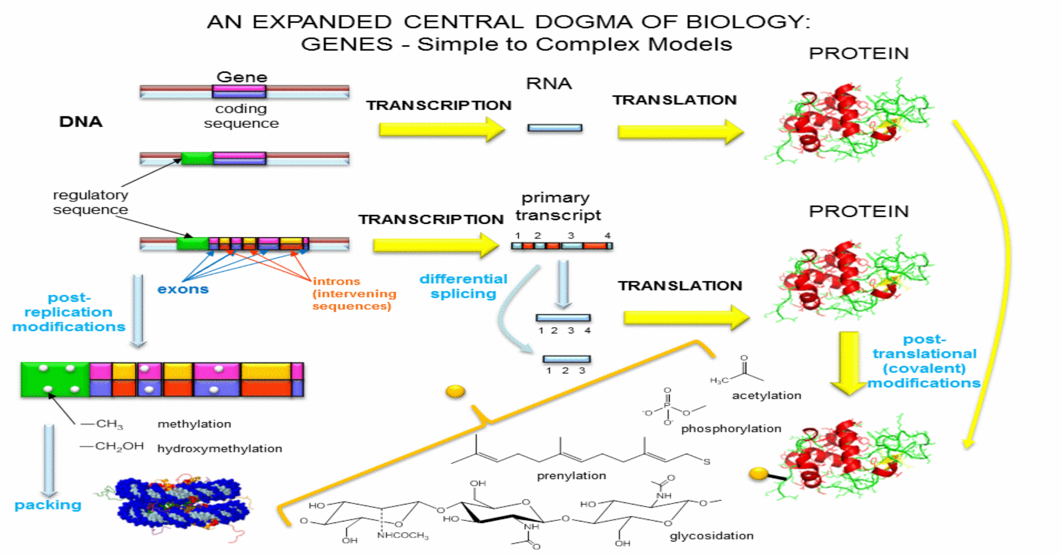


Figure 1.1. Schematic illustration of expended central dogma.
(Source: Jakubowski, 2011)

The term proteome was first introduced by Marc Wilkins in 1994. Wilkins abbreviated proteome from “PROTEin complement of the genOME” meaning the complete set of the proteins expressed by the cell (Wilkins et al., 1996). Then, the term proteomics was described as simply as “the large-scale analysis of proteins” (Pandey & Mann, 2000). Since mRNA transcripts can alternatively be spliced and translated proteins can be chemically modified in the cell, genomic studies alone cannot answer many questions which can be solved by proteomic approaches (Dunham et al., 1999), (Krogh, 1998).

Proteomics is the study of proteins on a large scale to obtain an integrated view of the cellular processes, networks at the protein level. Proteomic approaches contain sequence analysis, structure determination and functional exploration of proteins, as well as the study of protein-protein interaction. In order to explore a novel protein, sequence analysis is the first and indispensable step that provides information for structural and functional proteomics studies (Washburn, Wolters, & Yates, 2001). Edman degradation method is used for sequence analysis (Edman & Begg, 1967). In the identification of large amounts of proteins for sequence, this method is limited in that it takes a long time for chemical reaction to occur. In that sense, the method is not efficient enough. Nowadays, mass spectrometry replaces the classic Edman degradation method in protein and peptide sequencing due to its high accuracy, sensitivity and high throughput (Chait, 2006).

1.1.1. Mass Spectrometry (MS)

Mass spectrometry is an analytical technique to measure the mass-to-charge ratios (m/z) of ions. Although mass spectrometry has been used in analytical chemistry for over a century, only in the last two decades has the technology developed to offer adequately high resolution to make mass spectrometry an important tool for proteomics studies (Aebersold & Mann, 2003). A mass spectrometer has three main parts: an ion source that ionizes the molecules in the sample, a mass analyzer that separates the ions according to their mass-to-charge ratio, and a detector system that measures the abundance of ions with different mass-to-charge ratios. The main components of MS are illustrated in Figure 1.2.

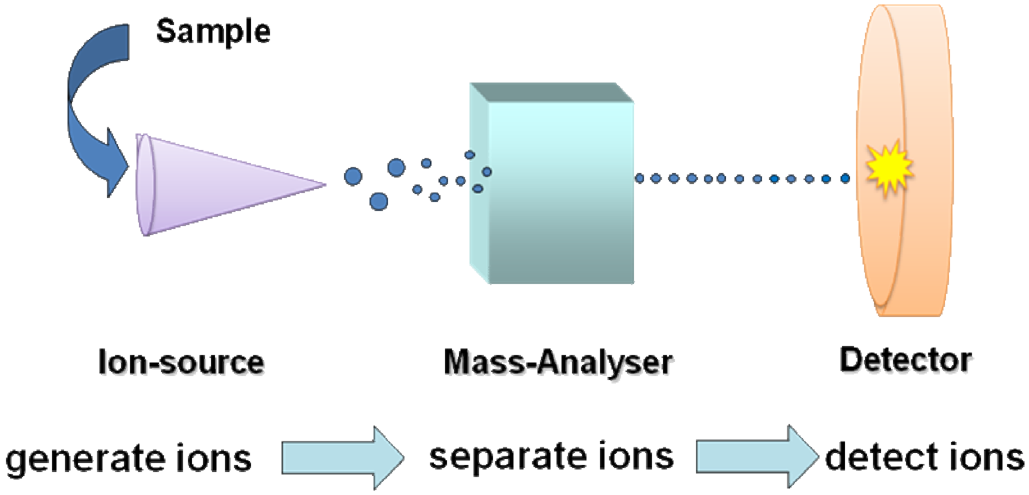


Figure 1.2. Main components of Mass Spectrometry. The figure adapted from Nathan Edward's slide. (Source: Edwards, 2005)

Mainly, mass spectrometry is used in protein identification. As shown in Figure 1.3, when only the first stage MS is used, the peptide fragments are gained. This is called peptide mass fingerprinting which can be used to identify the protein (Henzel et al., 1999). To identify a protein, at first, the protein or protein mixture of interest is digested by a protease. Then, the resulting peptide fragments are ionized and measured to generate a mass spectrum with high specificity. This machine-generated mass spectrum is named experimental spectrum. For each protein in a database, a theoretical mass spectrum can be constructed by using a virtual digestion of a protein. Protein identification is achieved by matching the monitored peptide masses in the experimental mass spectrum with the theoretical masses in the virtual mass spectrum gained from a sequence database (Hernandez, Müller, & Appel, 2006).

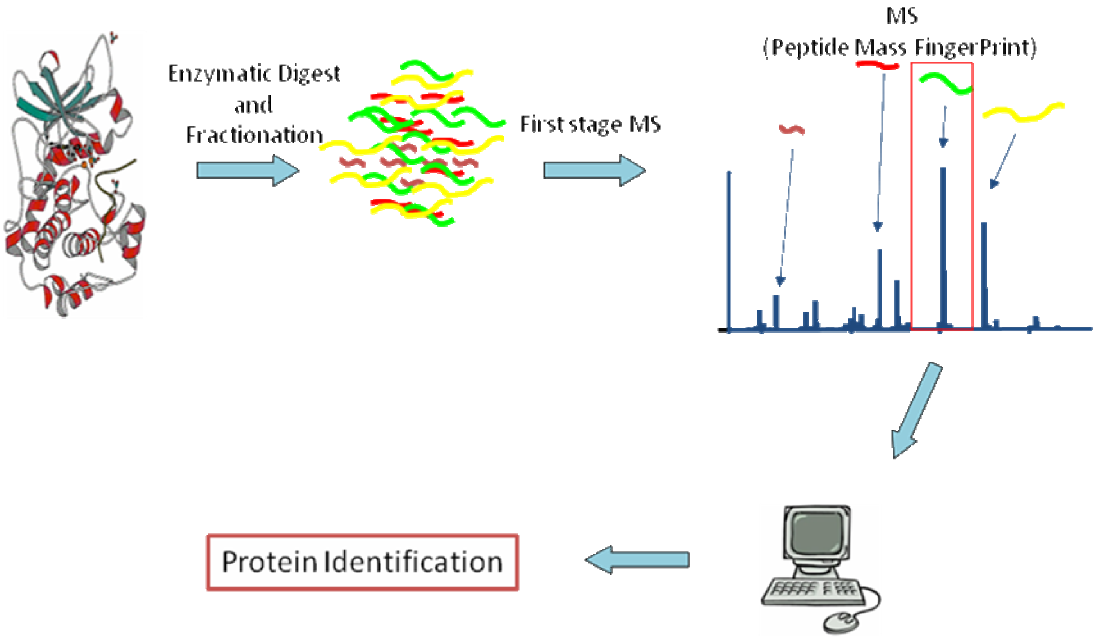


Figure 1.3. Protein identification by peptide mass fingerprint. The figure was adapted from Nathan Edwards’s slide (Source: Edwards, 2005).

However, this method suffers from the deficiencies that limit its utility in large proteomic studies. If the analyzed sample is very complex or the sequence database is very large, the protein assignments are likely to have low statistical significance. Tandem mass spectrometry becomes useful for high-throughput proteomic studies because it has a second stage of MS which gives more information about the peptide sequence (Bandeira, 2007).

1.1.1.1. Tandem Mass Spectrometry (MS/MS)

MS/MS is the most dominantly used tool in peptide identification and sequencing. Peptides are determined in complex mixtures by tandem mass spectrometry. Tandem mass spectrometry uses two mass analyzers in series to gain more detailed information about the protein of interest. A chamber where peptides are fragmented into small charged ions by Collision Induced Dissociation (CID) which is one of the several types of fragmentation methods connects the two mass analyzers. In Figure 1.4; MS1, MS2 and fragmentation region (chamber) are illustrated. The first stage MS (MS1) functions as a normal mass analyzer, where peptide fragments of a protein or protein mixture are measured according to their mass-to-charge ratios. Although all peptides go through the connection chamber, only peptides which are within a small range of m/z ratios are selected for CID. With inert gas atoms in the chamber the ions are collided and then the emerging ion fragments get in the second mass analyzer (MS2). Subsequently their m/z ratios are measured to generate a tandem mass spectrum (Kinter & Sherman, 2000).

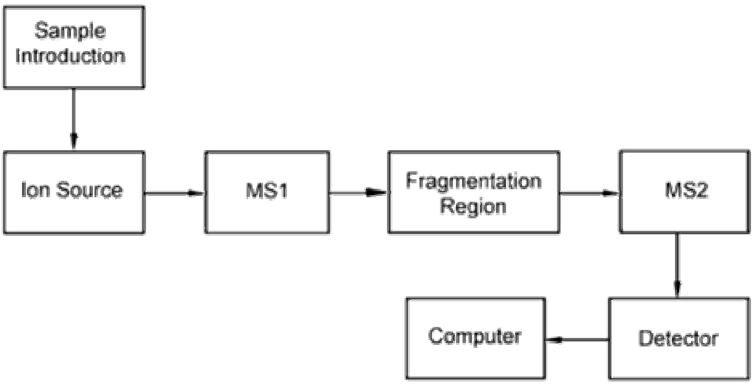


Figure 1.4. Components of Tandem Mass Spectrometry (MS/MS).
(Source: Zaidi, 2006)

Tandem mass spectrometry is mainly used in peptide identification. In order to identify a peptide, proteins can be digested theoretically and the virtual spectra can be compared to the experimental spectrum. This method is called database search. Also, the sequence information of the peptide can be identified by using the experimental spectrum directly without need for the protein database. This method is called *de novo* peptide sequencing. Detailed information about database search and *de novo* sequencing will be given in sections 1.1.2.1 and 1.1.2.2 After peptide identification is performed, proteins can be identified (Figure 1.5). In order to find a peptide sequence from a tandem mass spectrum, understanding the peptide fragmentation processes is an important point.

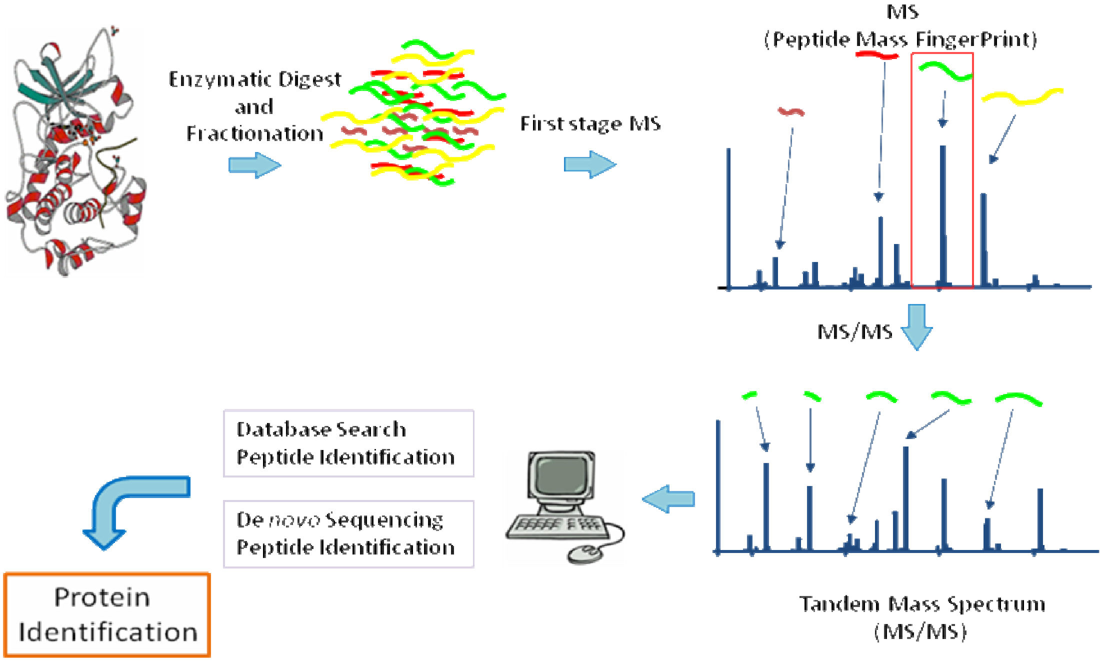


Figure 1.5. Protein identification by tandem mass spectrum. The figure was adapted from Nathan Edwards's slide (Source: Edwards, 2005).

1.1.1.2. Peptide Fragmentation

The most common method of fragmentation is CID (Wells & McLuckey, 2005). In the CID fragmentation process, for instance a peptide bond at a peptide molecule can break. The peptide is fragmented into two ions which are either from the N-terminus or from the C-terminus. In addition to b- and y-ions, in MS/MS spectra various types of ions can appear due to a variety of reasons. The first reason is "neutral loss ions". The b- and y-ions may lose certain chemical groups such as H_2O and NH_3 and as a result neutral loss ions are formed (Papayannopoulos, 1995). The second reason is the "internal fragments". The fragmentation can also occur at other positions in the backbone. If it occurs between the two carbons, the resulted ions are called a-ion and x-ion. If it occurs between a nitrogen and a carbon, the resulted ions are called c-ion and z-ion. As a result, the internal fragments are formed when more than one bond of peptide is broken. The third reason is "doubly or multiple charged b- and y-ions". When more than one proton exists on the fragments, multiple charged b-ions and y-ions occur. The last reason is "unexplained data" of the spectrum. The pattern of occurrence of various ions is called fragmentation pattern of a peptide. The ions represent a characteristic pattern in tandem mass spectrometry which allows identification of selected peptides. However, actual tandem mass spectra are much more complicated due to occurrence of unknown ion types, unknown charges, missing ions, unexplained data, isotopic ions, and machine errors. Therefore, the identification of peptide sequences using tandem mass spectra remains a challenging task (Huang et al., 2005).

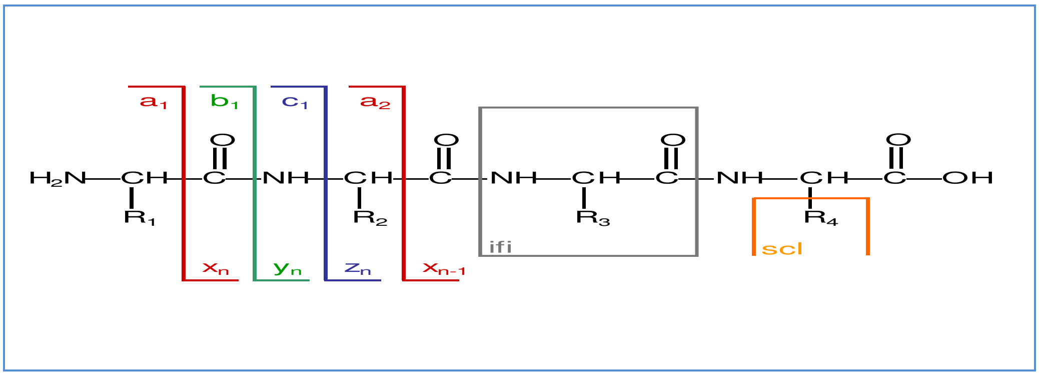


Figure 1.6. Fragmentation patterns of a peptide containing *n* residues
(Source: Allmer, 2006).

Possible ions resulting from peptide fragmentation are illustrated in Figure 1.6 When the peptide bond, which is between two amino acids, is broken, b_1 and y_n ions are generated. These ions are complementary ion pairs because the combination of them gives the complete peptide sequence. From a peptide with *n* amino acids, *n*-1 b-ions and *n*-1 y-ions are generated by an ideal fragmentation. A complete b-ions and y-ions ladder, where two adjacent m/z values differ by fragment mass of only one amino acid, are formed by these b- and y-ions form. Table 1.1 illustrates an example of the b and y ion ladders. If the tandem mass spectrum does not include any unexplained data and contains only b-ion and y- ion types, it is much easier to find a complete b-ion or y-ion ladder.

Table 1.1. Ionization and fragmentation of peptide NH_3 R1 - R2 - R3 - R4 COOH. R (radical) group differs for each amino acid.

Ions	b-ion sequences	Ions	y-ion sequences
b1	$(R1)^+$	y3	$(R2 - R3- R4)^+$
b2	$(R1 - R2)^+$	y2	$(R3 - R4)^+$
b3	$(R1 - R2 - R3)^+$	y1	$(R4)^+$

1.1.2. MS/MS-Based Peptide Identification

In order to find a peptide sequence from a MS/MS spectrum, generally two approaches are used: database searching and *de novo* sequencing. The database search method compares experimental MS/MS spectra with theoretically generated MS/MS spectra of each peptide derived from a protein sequence database, assuming that the query peptides exist in the protein sequence database. On the other hand, *de novo* sequencing method assigns amino acid sequences to MS/MS spectra without a need for a protein database (Forner, Foster, & Toppo, 2007).

1.1.2.1. Database Search

In database search methods, at first the proteins in the database are virtually digested by the protease. According to the molecular weight of the resulting peptides, they are then indexed. Only those with an approximate molecular weight matching to the precursor mass of the tandem mass spectrum are considered as candidates. Then, the theoretical spectra of those candidates are generated. In order to calculate the scores for the match of the theoretical spectrum and experimental spectrum, a scoring function is applied. The candidate which has the highest match score is close to be the peptide which generated the experimental spectrum. The database search method is illustrated in Figure 1.7 Since it depends on the protein database, the size of the database influences the searching speed and the prediction of the results. For instance, although a larger database provides more results, it requires longer searching time and leads to large false positive rates. On the contrary, smaller database can search faster but can also result in false peptide prediction.

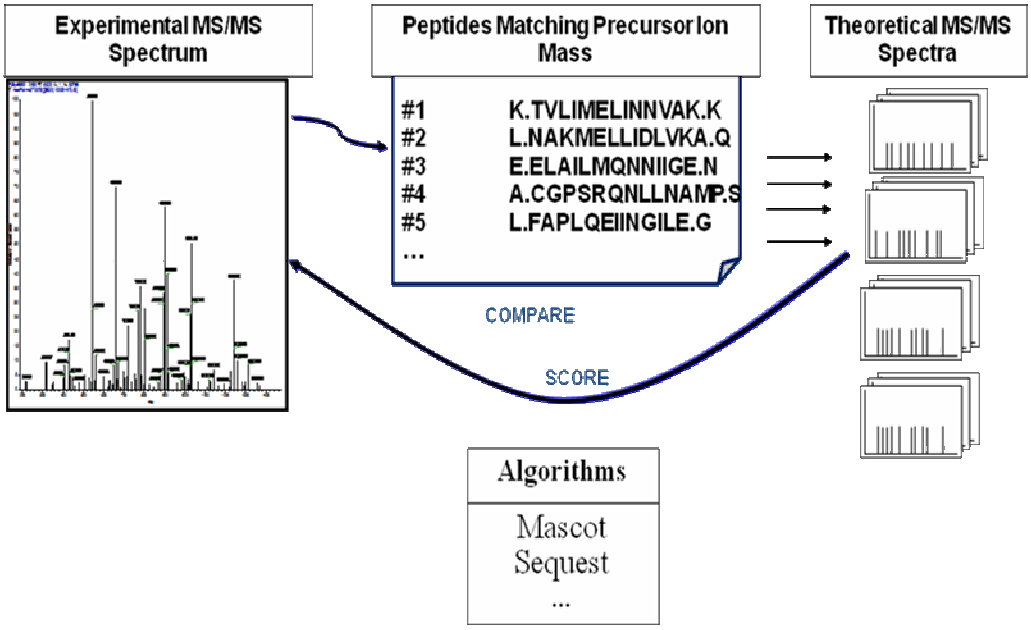


Figure 1.7. Database Search Method. Precursor mass of MS/MS spectrum can match so many peptide sequences from database. After generating theoretical MS/MS spectra from those candidates, match score is calculated for each candidate. The candidate which has highest match score can be the peptide sequence of experimental MS/MS spectrum.

The most popular database search programs include SEQUEST (Eng, McCormack, & Yates, 1994), Mascot (Perkins, Pappin, Creasy, & Cottrell, 1999) and OMSSA (Godoy, Olsen, Souza, Li, Mortensen, & Mann, 2006). The database search algorithm is effective and has been used successfully in proteomics projects for organisms with well-studied genomes. However, it is not applicable in situations where a target sequence does not exist in the protein database. If proteins under study are novel, carry unknown mutations or post-translational modifications, a target sequence does not exist in the database. Because of the disadvantage of the database searching method, a variety research has focused on *de novo* sequencing methods.

1.1.2.2. *De Novo* Sequencing

The aim of *de novo* sequencing algorithms is to reconstruct the peptide sequence directly from a tandem mass spectrum without the aid of databases. Therefore, in principle, the *de novo* sequencing method can cope with limitations of the database search algorithms. Various *de novo* sequencing algorithms have been developed since the late 1990's, including Sherenga (Dancik, Addona, Clauser, Vath, & Pevzner, 1999), Lutefisk (Taylor & Johnson, 1997), the dynamic programming method (Chen, Kao, Tepel, Rush, & Church, 2001) and a suboptimal method (Lu & Chen, 2003), PEAKS (Ma, Doherty-Kirby, & Lajoie, 2003), DACSIM (Zhang Z., 2004), EigenMS (Bern & Goldberg, 2005), PepNovo (Frank & Pevzner, 2005) (Frank, Savitski, Nielsen, Zubarev, & Pevzner, 2007) and NovoHMM (Fischer et al., 2005). The basic problem of these algorithms is that they are inadequate to obtain a complete sequence or a full ladder of b- or y-ions. The mass difference between consecutive steps in a ladder or peaks is equal to an amino acid mass. Therefore, full ladder enable to find amino acid sequence of a peptide. This process is illustrated in Figure 1.8 However, in practice, due to incomplete fragmentations, isotopes, multiple fragmentations, unknown ion type and random unexplained data in tandem mass spectrum seriously restricts the efficacy of *de novo* algorithms stated above, and frequently causes false positive predictions. Hence, nowadays, *de novo* sequencing tools are not as extensively used as database search tools.

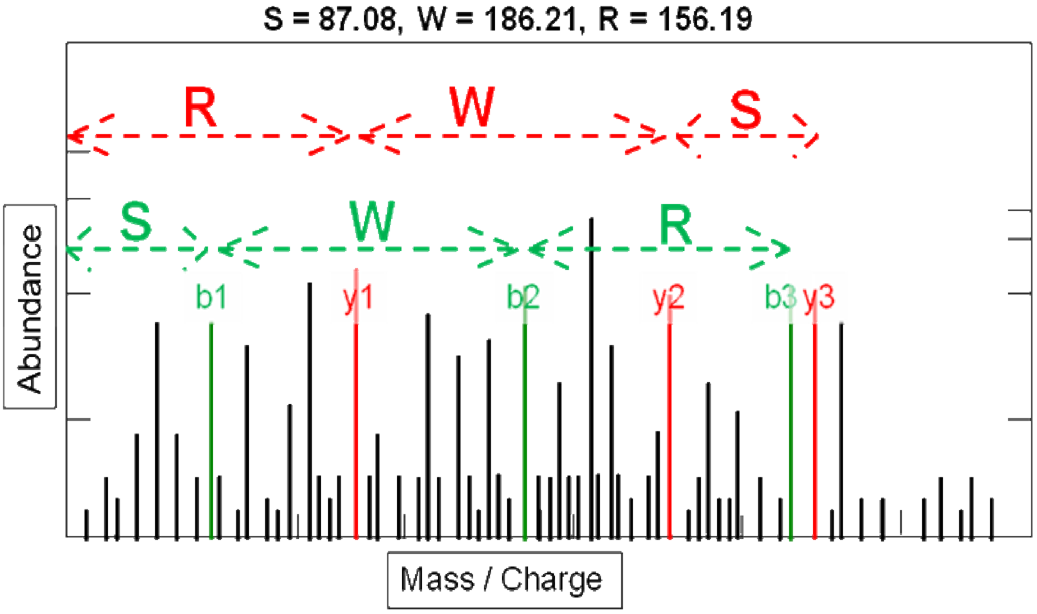


Figure 1.8. *De novo* sequencing Method. Full b-ion and y-ion ladders are shown. Mass difference between b-ion or y-ion ladder gives the peptide sequence of a tandem mass spectrum (Source: Chen, Kao, Tepel, Rush, & Church, 2001).

In order to develop a reliable *de novo* sequencing algorithm, there are two main steps. The first one is generating a pool of candidates (Eng, McCormack, & Yates, 1994). To generate a pool of candidates, most of the *de novo* sequencing algorithms create a spectrum graph. The second one is designing a good scoring function to choose the best candidate from the pool (Havilio & Smilansky, 2003). In order to choose the best candidate, a dynamic programming algorithm is mostly used to find paths in the graph which represent peptide candidates.

The dynamic programming algorithm works only if the scoring function is additive. Therefore, the selection of scoring functions is quite limited (The sum or the multiplication of the score of each ion equals the score of a peptide). Despite this limitation, various scoring functions have been proposed. The probability of different ion types in their scoring function is considered by Dancik (1999). Later, Frank and Pevzner (2005) developed a *de novo* sequencing algorithm which is called PepNovo by using a probability network with a hypothesis testing. At the same time, Bernd Fischer et al. (2005) proposed an HMM model called NovoHMM. In recent years, Frank and Pevzner (2005) presented a new version of PepNovo which suits to LTQ-FT data (Frank, Savitski, Nielsen, Zubarev, & Pevzner, 2007). The new version of PepNovo creates sequence tags up to 8 amino acids long which are then used to query the database for peptide identification and is therefore not a pure *de novo* sequencing algorithm.

1.1.3. *De Novo* Sequencing Algorithms

Over the years, in order to deal with the *de novo* sequencing problem, various algorithms have been developed. One of the early *de novo* sequencing algorithm uses the “naive method” which was developed by Sakurai, Matsuo, Matsuda and Katakuse (1984) and Hamm, Wison and Harvan (1986). This method lists all the possible candidate peptides according to the mass of the precursor ion of the tandem mass spectrum. Therefore, it is also called “exhaustive listing”. In order to find the best candidate peptide with the highest match, all the candidate peptides are compared with the experimental spectrum (tandem mass spectrum). However, there is a computational hardness because possible enormous list of candidates will exist for a usual precursor mass (Lu & Chen, 2004).

The second method is called “subsequencing” which was developed by Johnson and Biemann (1989). In this approach, short sequences represent only a part of the whole sequence. These short sequences (subsequences) account for some observed ions. The subsequences are extended to one residue until the whole sequence is tested. During this extension, if there is a significant matching with the experimental spectrum, those subsequences will be retained. However, the method is problematic in that some good candidate peptides might be discarded when some regions of a peptide are inadequately represented by fragmentation ions. It is important to highlight that the fragmentation frequencies are usually not evenly distributed over the whole peptide (Lu & Chen, 2004).

The third method uses graphical display of the data which was developed by Scoble, Billet and Biemann (1987). In this approach, the same ion type series are allowed to find the peptide sequence because connections occur if the mass differences between the same type fragmentation ions equal to one of the amino acid mass. The approach is limited in high-throughput environments. However, it can be used for manual *de novo* interpretation of tandem mass spectra (Lu & Chen, 2004).

The fourth method uses "graph theory" which was first proposed by Bartels (1990). Later, the method was further developed by Hines, Falick, Burlingame and Gibson (1999), Fer nandez-de-Cossio, Gonzales and Besada (1995), Taylor and Johnson (1997) and Dancik, Addona, Clauser, Vath and Pevzner (1999). In this method, the experimental spectrum is transformed into a graph called "Spectrum Graph". Each peak in the experimental spectrum corresponds to a vertex in the spectrum graph, and a directed edge is drawn between two vertices if the mass difference of the two vertices equals the mass of one amino acid mass (Lu & Chen, 2004). In order to find paths, one of which can represent the peptide sequence of experimental spectrum, various algorithms have been designed, (Figure 1.9).

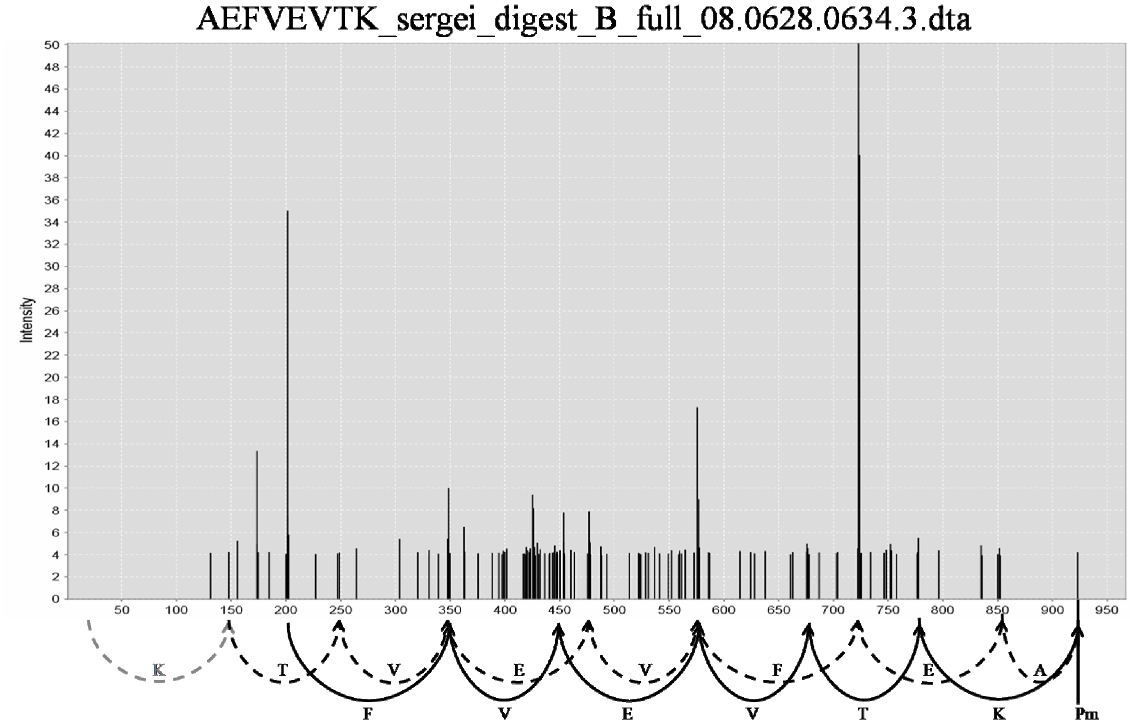


Figure 1.9. Spectrum Graph
(Source: Allmer, 2011).

Nowadays, for peptide identification, the most popular *de novo* sequencing algorithms are PEAKS, PepNovo, Lutefisk, Sherenga and some of which use the spectrum graph approach. In the next section, in addition to these, the Dynamic Programming and the Suboptimal Method are explained in more detail.

1.1.3.1. PEAKS

Ma, Doherty and Lajoie (2003) developed PEAKS. In this method, the raw MS/MS data is preprocessed by filtering unexplained data, centering peaks and deconvolving doubly and triply charged peaks into singly charged peaks. Based on how close a peak matches to a mass in the theoretical spectrum, reward or penalty is given. Ma et al (2003) have aimed to find sequence whose b- and y-ions maximizes the rewards at their mass value. Dynamic programming has been used to compute the 10000 sequences with the highest scores. After that, the scores are refined by a more stringent scoring scheme which uses stricter mass error tolerance and then the best candidates are found using the new scoring scheme. In order to calibrate the minor deviation in the MS/MS data, a recalibration method has been used. Finally, PEAKS calculates a confidence score for each top-scoring peptide sequence (Pevtsov, Fedulova, Mirzaei, Buck, & Zhang, 2006).

1.1.3.2. PepNovo

The PepNovo program which was developed by Frank and Pevzner (2005) is based on a probabilistic network modeling scoring method. A probabilistic network has been defined by three different types of dependencies: correlations between fragment ions, dependencies because of the relative position of the cleavage site in the peptide, and the influence of flanking amino acids to the cleavage site. Frank and Pevzner (2005) has integrated a hypothesis test idea into their scoring function and defined the score given to a mass m and spectrum S to be the logarithm of the likelihood ratio of the probabilities of the CID and random hypotheses. While the CID hypothesis assumes that mass m is caused by a genuine cleavage in the peptide, the random hypothesis assumes that mass m is caused by a random process.

A spectral graph is built using only the top-ranking peaks in a tandem mass spectrum by a sliding window method. Also by using a combinatorial parent mass correction procedure, the parent mass is recalibrated. A dynamic programming which is similar to Chen et al's algorithm is applied to find the highest asymmetric path in the spectrum graph. The uniqueness of Pepnovo comes from its scoring function (Bern, Cai, & Goldberg, 2007).

1.1.3.3. Lutefisk

The Lutefisk algorithm, which was developed by Taylor and Johnson (1997) identifies significant ions in the first step. In this step, local maxima and sliding windows methods are used. Then, the N- and the C- terminal evidence lists are detected. While the N-terminal ions are considered as b, b-NH_3, b-H_2O, a, a- NH_3, a- H_2O, the C-terminal ions are considered as y, y- NH_3 and y- H_2O. Taylor and Johnson (2001) used an approximate probability the each ion, those of a, a- NH_3, a- H_2O, b- NH_3, b- H_2O, y- NH_3 and y- H_2O are half of that assigned to b and y. In the next step, the N-terminal and C-terminal evidence lists are combined into a spectrum graph in which the x-ordinate is the nominal m/z values for b-ions and the y-ordinate is a sum of the various ion probabilities suggesting cleavage at each site. Once the spectrum graph is established, the program proceeds by tracing out sequences starting from the N-terminal. The process starts by finding b-ions which differ from N-terminal by one or two amino acids residue mass. After getting the completed sequences, b-ion values are scored according to an intensity-based score (Taylor & Johnson, 2001).

1.1.3.4. Sherenga

Developed by Dancik, Addona, Clauser, Vath and Pevzner (1999), Sherenga is the first algorithm importing the concept of spectrum graph and training data. In this algorithm, a set of training data is used to learn ion types of the peaks in the spectrum without any prior information from the fragmentation patterns. Then, in order to build a spectrum graph, the information of the ion types is used. Different ion types have been defined as a set, k is the number of ion types. Each peak in the spectrum creates k vertices. Two vertices u and v are connected by directed edge only if the mass difference between them equals to one amino acid. Dancik et al (1999) indicated that the peptide sequencing problem is transformed as the longest path problem in the directed acylic graph and that the longest path may not correspond to exact solutions because it may use multiple vertices associated with the same experimental peak. Finding the longest non-symmetric path is one solution to avoid multiple assignments of the same peaks. Dancik et al (1999) indicated that to find the longest non-symmetric path in the spectrum graph there is a need for the existence of an efficient algorithm.

1.1.3.5. Dynamic Programming

The first application of dynamic programming in *de novo* sequencing was proposed by Chen, Kao, Tepel, Rush and Church (2001). They first found forbidden pairs in the spectrum graph were non-interleaving. Then, in order to find the longest non-symmetric path in the spectrum graph, a dynamic programming method which makes the first polynomial time algorithm for *de novo* sequencing was designed. When Dancik's idea (1999) is simplified, each peak is either b or y ions. Then, instead of the k vertices, each peak generates two vertices. If their mass difference equals to an amino acid, two vertices are connected. Later a dynamic programming which is a common technique for solving optimization problems is used to get the optimal results. However, the optimal solution may not be the correct sequence which is produced by the experimental spectrum because unexplained data and unknown ions are interpretated as real ions. For this reason, scientists are also interested in suboptimal solutions to reach the real sequence (Goto & Schwabe, 2008).

1.1.3.6. Suboptimal Methods

The suboptimal concept which was developed by Lu and Chen (2003) extended the previous dynamic programming method. In this method, the spectrum is transformed into a two dimensional matrix spectrum graph and then the suboptimal solutions are found. For each candidate peptide, a theoretical spectrum is generated and scored by using a simple scoring function. While S1 indicates the sum of the abundance levels of all of the ions in the theoretical spectrum, S2 represents the sum of the abundance levels of the ions that match with the experimental spectrum. In order to rank the results, the ratio of S2/S1 is used as the score. The ratio of S2/S1 shows how well each theoretical spectrum suits the experimental spectrum. Therefore, the candidate peptides are ranked according to the S2/S1 ratio.

1.2. Aim of the Study

The amino acid sequence of a protein has to be determined in order to solve its structure and function. Since the identification of the sequence of a peptide from the tandem mass spectrum is the first and common step to identify the protein, correct peptide identification is crucial for accurate protein sequence determination.

Tandem mass spectrometry, which is a powerful tool for peptide sequencing, measures the mass to charge ratio of each peptide fragment. Since ion fragments information is very complex and can often lead to incorrect identification, diverse *de novo* sequencing and database search algorithms have been developed to gain correct peptide sequence.

Although database searching is usually the first choice for peptide identification, *de novo* peptide sequencing plays an important role in dealing with the limitation of database searching. Even though it is crucial to identify peptide sequences accurately in protein sequencing, unfortunately current *de novo* peptide sequencing algorithms cannot determine the peptide assignment with high correctness. Therefore, finding an accurate peptide sequence is still a challenging task.

The aim of this study is to enhance the accuracy of the currently used *de novo* sequencing algorithms. In order to manage this goal, the present study focuses on the problems such as missing ions from tandem mass spectrum, unexplained data and unknown ion types. By using ion complementarity, the fragment ions are named. Therefore, named ions provide determination of missing ions which are crucial to obtain the b-ion ladder. Finding the b-ion ladder provides to get rid of unexplained data which causes complexity to tandem mass spectrum. Then, in order to determine the peptide sequence of tandem mass spectrum, the b-ion ladder is traversed using dynamic programming to derive the amino acid sequence.

CHAPTER 2

MATERIALS AND METHODS

2.1. Materials

2.1.1. Tandem Mass Spectra Data Set

In order to evaluate quality of the *de novo* sequencing algorithm, one public data set which was developed and published by Keller et al (2002) and which was generated with a LCQ tandem mass spectrometer was used. One feature of the data set of low energy CID tandem mass spectra is that it was generated from a control mixture of known protein components. Keller and colleagues searched the human peptide database to find the correct peptide assignment with SEQUEST. Since they did not gain all of the peptide assignments accurate, alternative database search methods were used to find correct peptide identification in the rest of the spectra (Keller et al, 2002).

The data set includes peptide assignments for Spectra of Singly Charged Ions, Spectra of Doubly Charged Ions and Spectra of Triply Charged Ions. They were used in the several studies to determine the quality of *de novo* sequencing algorithm.

2.1.2. Computer

Another basic material used in the study is a computer which was used to keep the data set. In addition, NetBeans IDE platform was downloaded onto the computer, in order to generate the *de novo* sequencing algorithm by using Java programming language. After the algorithm was developed, it was tested on the data set to evaluate the algorithm.

2.1.3. Java

Java, which is a programming language, was developed by Oak. It was designed for handheld devices and set-top boxes. Since Oak was not successful, in 1995 Sun altered the language to take the advantage of the burgeoning World Wide Web. Also, Oak's name was changed to Java (Bilsel, 2009).

Java has significant features such as Platform Independence, Object Orientation, Rich Standard Library, Applet Interface and Garbage Collection. Java compilers do not generate native code for a particular platform but rather 'byte code' instructions for the Java Virtual Machine (JVM). This feature of java is referred as Platform Independence. Making Java code operate on a specific platform is then simply a matter of writing a byte code interpreter to simulate a JVM (Figure 2.1). What this all means is that the same compiled byte code will run unmodified on any platform that supports Java (Daconta, 1996). Java is a pure object-oriented language. This property of java is called Object Orientation. It means that everything in a Java program is an object and everything is descended from a root object class. One of Java's most attractive features is its standard library which is called Rich Standard Library.

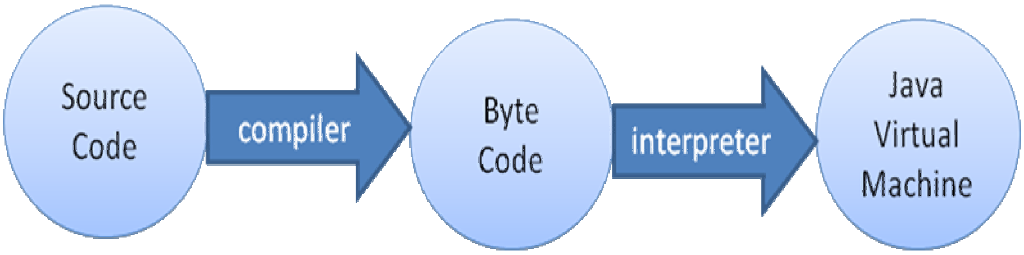


Figure 2.1. Java Execution Model. A Java Virtual Machine is a virtual machine capable of executing Java Byte Code which is compiled from Java Source Code.

The Java environment includes hundreds of classes and methods in six major functional areas (Chan & Lee, 1996). The first one is Language Support classes for advanced language features such as strings, arrays, threads, and exception handling. The second one is Utility classes like a random number generator, date and time functions, and container classes. The third one is Input/output classes to read and write data of many types to and from a variety of sources. The fourth one is Networking classes to allow inter-computer communications over a local network or the Internet. Abstract Window Toolkit for creating platform-independent GUI applications (Flanagan, 1996). In addition to being able to create stand-alone applications, Java developers can create applets that can be downloaded from a web page and run in a client browser (Gosling & Yellin, 1996). Also, Java does not require programmers to explicitly allocated memory. This feature is called Garbage Collection which makes Java programs easier to write and less prone to memory errors (Lemay & Perkins, 1996). Additionally, the popular C++ programming language is similar to the Java Syntax. Hence, this similarity is one of the factors enabling the rapid adoption of Java (Niemeyer & Peck, 1996).

2.1.4. Netbeans

NetBeans is both a framework for Java desktop applications and an integrated development environment (IDE). The Netbeans Platform provides a reusable framework in order to simplify the development of Java Swing (Swing is the primary Java GUI widget toolkit) desktop applications and offers reusable services such as user interface management (for instance menus and toolsbars), user setting management, storage management, window management, wizard framework, netbeans visual Library, integrated development tools.

An IDE is computer software to help computer programmers develop software such as Java, JavaScript, PHP, Phthon, C, C++. The NetBeans IDE is written in Java. It can be run in Windows, Linux, Solaris and Mac OS, if the JVM is installed. For developing Java, a Java Development Kit (JDK) which is a Sun Microsystems product aimed at Java developers is necessary (Tulach, 2010).

2.1.5. TortoiseSVN

TortoiseSVN is a free open-source client for the Subversion version control system. Version control is a critical tool for programmers because it enables them to manage changes of files. The files are stored in a central repository. The repository can remember every change ever made to the files. Therefore, older versions of the files can be recovered. In addition, the client can reach all the information about the files such as how and when the changes are done, and who changed them (Large, Küng, & Onken, 2011).

The features of TortoiseSVN are shell integration, icon overlays and easy access to Subversion commands. Shell integration means that TortoiseSVN integrates free of problems into the Windows shell (i.e. the explorer). Therefore, working with the tools which are familiar to the user can be kept. Icon overlays indicate that the status of every versioned file and folder. Hence, it is possible to see instantaneously what the status of working copy is. Easy access to Subversion commands means that all Subversion commands are available from the explorer context menu and Submenu of TortoiseSVN is added to there (Küng, Ongen, & Large, 2011).

In the present study, in order to reach older versions of Java code which includes implementation of the algorithm, Subversion version control system was used in Origo.

2.1.6. Origo

Origo, which is a hosting platform, is available for both open-source and closed-source projects. It assures the required mechanisms of a project that are an SVN repository, a template project wiki page, an issue tracker with tag support and a place to put releases (Origo, 2007).

2.1.7. MS Product

MS Product program which is one of the Protein Prospector Tools was developed in the University of California, San Francisco (UCFS). The Mass Spectrometry Facility is conducted at UCSF by Dr. Alma Burlingame, Professor of Chemistry and Pharmaceutical Chemistry (Burlingame). MS Product calculates theoretical ion masses from given peptides. The peptides undergo dissociation via post-source decay or high- or low-energy collision-induced dissociation (Agilent Spectrum Mill MS Proteomics Workbench, 2010).

Since MS Product presents the ion fragments with correct name from the given peptide, the accuracy of the ion naming algorithm was ensured by testing with MS Product. Peptide fragmentation model was used to build the name of the fragment ions.

2.1.8. Peptide Fragmentation Pattern

In order to name ions in tandem mass spectra, the peptide fragmentation model was used as indicated in Figure 2.2 The model enables us to exact mass differences between fragment ion types which are crucial for the calculation of the complements of the fragment ions.

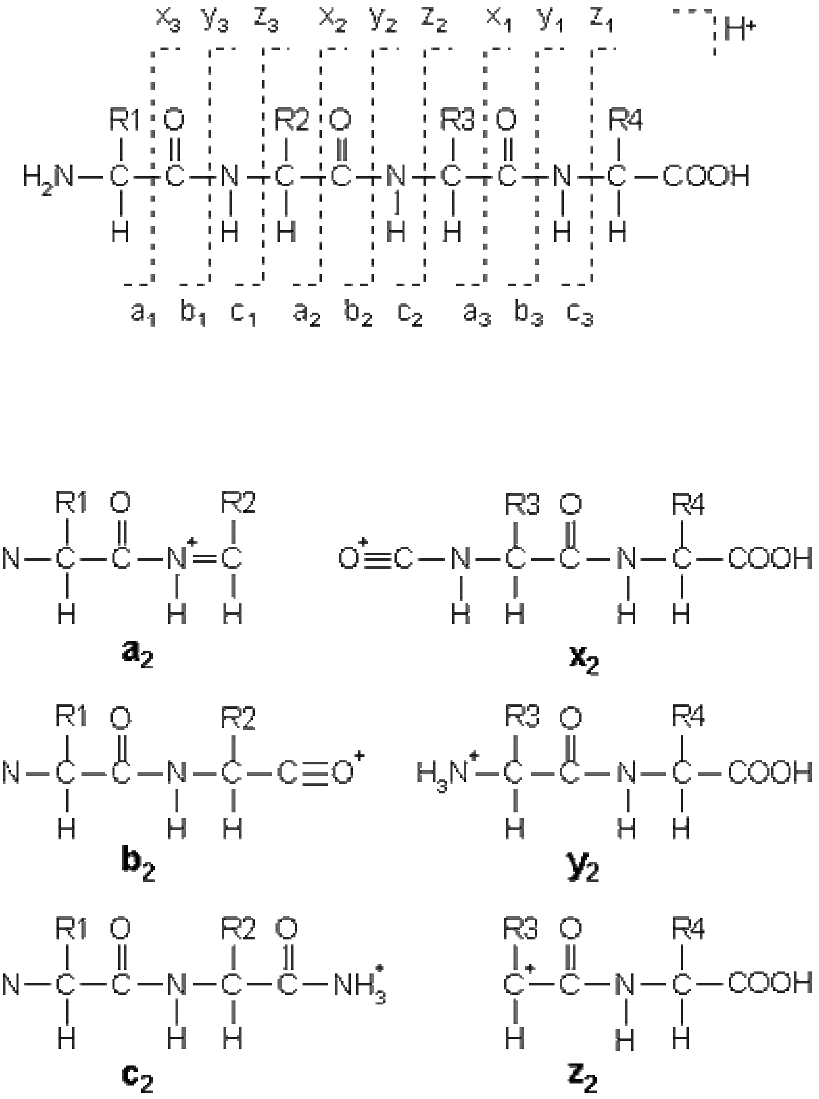


Figure 2.2. Ion Complementarity by Peptide Fragmentation Model. Fragmentation regions and formulas of backbone fragment ions are shown (Source: Peptide fragmentation, 2010).

2.1. 9. Amino acid Mass

The presence of isotopes (C=12, C=13, C=14 etc.) at their natural abundances makes it essential to define whether an experimental mass value is an "average", or a "monoisotopic" value, While the monoisotopic mass is the mass of the first peak of the isotopic distribution, the average mass is the average masses of all of the peaks in the isotopic distribution.

In order to clarify this situation, an example could be given. For instance, the monoisotoptic mass of a compound is 1155.6. For a given compound the monoisotopic mass is the mass of the isotopic peak whose elemental composition is composed of the most abundant isotopes of those elements. The monoisotopic mass can be calculated using the atomic masses of the isotopes. The average mass is the weighted average of the isotopic masses weighted by the isotopic abundances. The average mass can be calculated using the atomic weights of the elements (Figure 2.3) (Ion source, 2005).

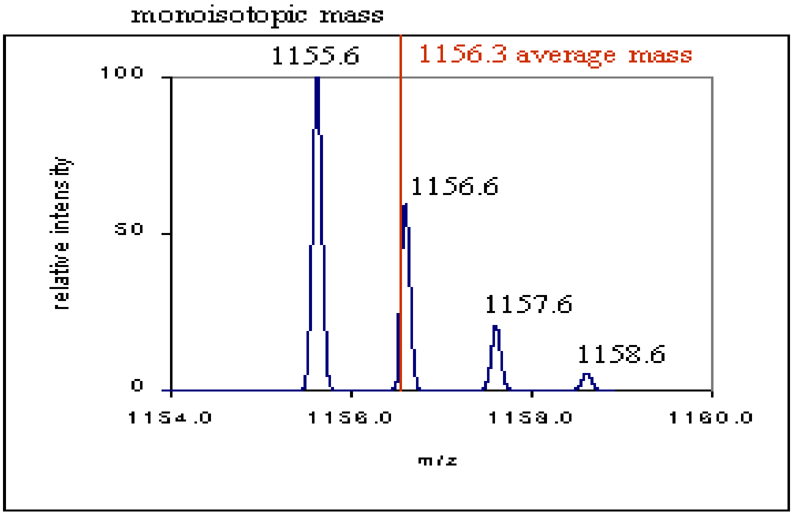


Figure 2.3 Monoisotopic and Average Mass
(Source: Ion source, 2005).

In the present study, both monoisotopic and average mass of the amino acid was considered for the tandem spectra dataset. In order to generate theoretical precursor ion mass from given peptide sequence of the dataset, the monoisotopic masses of the amino acids were used (Figure 2.4). In addition to the calculation of the theoretical precursor ion mass, the amino acid masses were used to generate theoretical ions masses.

Additionally, in the present study, the amino acid masses were used for determination amino acid sequence of the peptide. According to the value of the precursor ion mass, monoisotopic or average mass of amino acids were applied to find the peptide sequence.

Table 2.1. Monoisotopic and Average mass of each one of the amino acids.

1-letter code	3-letter code	Chemical formula	Monoisotopic	Average
A	Ala	C_3H_5ON	7.103.711	710.788
R	Arg	$C_6H_{12}ON_4$	15.610.111	1.561.875
N	Asn	$C_4H_6O_2N_2$	11.404.293	1.141.038
D	Asp	$C_4H_5O_3N$	11.502.694	1.150.886
C	Cys	C_3H_5ONS	10.300.919	1.031.388
E	Glu	$C_5H_7O_3N$	12.904.259	1.291.155
Q	Gln	$C_5H_8O_2N_2$	12.805.858	1.281.307
G	Gly	C_2H_3ON	5.702.146	570.519
H	His	$C_6H_7ON_3$	13.705.891	1.371.411
I	Ile	$C_6H_{11}ON$	11.308.406	1.131.594
L	Leu	$C_6H_{11}ON$	11.308.406	1.131.594
K	Lys	$C_6H_{12}ON_2$	12.809.496	1.281.741
M	Met	C_5H_9ONS	13.104.049	1.311.926
F	Phe	C_9H_9ON	14.706.841	1.471.766
P	Pro	C_5H_7ON	9.705.276	971.167
S	Ser	$C_3H_5O_2N$	8.703.203	870.782
T	Thr	$C_4H_7O_2N$	10.104.768	1.011.051
W	Trp	$C_{11}H_{10}ON_2$	18.607.931	1.862.132
Y	Tyr	$C_9H_9O_2N$	16.306.333	1.631.760
V	Val	C_5H_9ON	9.906.841	991.326

2.2 Method

2.2.1. Default Values

In the present study, proper default values were established. Default values were generated to reach the best settings for the Keller et al spectra data set. A Precursor mass tolerance, a fragment ion tolerance, ion types and windows settings were considered as default values to detect the best settings.

For each setting, we calculated the statistics for all spectra. In order to determine the best settings, the result of statistics was compared to the current best result. In other words, we checked whether the results for a new setting were better than the ones that were currently the best. Therefore, when the new one was better, we accepted it as the best setting and overwrote to the current best one. In contrast, when the new one was worse, it was ignored and the current best one was saved.

In the study, we detected the best settings to get statistics showing (found b- ion /theoretical-b-ion) score at maximum level, the minimum number of peaks in Named Spectrum and the minimum runtime. In accordance with this, we used those settings and similarly determined the best settings for b-ion Spectrum. In Chapter 3, the statistic is used to filter the named spectrum and b spectrum.

CHAPTER 3

DENOVON

3.1. Ion Complementarity

As indicated in sections 2.1.8 and 1.1.1.3, if the charge is retained on the N-terminus, peptide fragment ions are indicated by a, b, or c. On the contrary, if the charge is maintained on the C-terminus peptide fragment ions are stated by x, y or z. The subscript indicates the number of amino acid residues in the fragment. Superscripts are sometimes used to indicate neutral losses in addition to the backbone fragmentation, for the loss of ammonia and for loss of water. Although peptide backbone cleavage is the most useful for sequencing and peptide identification, other fragment ions may be observed in the tandem mass spectrum (Johnson, Martin, & Biemann, 1988) (Falick, Medzihradszky, Baldwin, & Gibson, 1993). Therefore, in the study not only b-type ions and y-type ions but also other ion types are considered to name ions by using ion complementarity (Table 3.1).

Table 3.1. Ion Complementarity between fragment ions. (O = 15.999, C = 12.011, N = 14.007, H = 1.0079).

	a	b	c	x	y	z
a		+C+O	+C+O+ N+3H			
b	-C-O		+N+3H			
c	-C-O- N- 3H	-N-3H				
x					-O-C+ 2H	-O-C- N-H)
y				+O+C-2N		-N-3H
z				O+C+N+ H	+N+3H	

We generate neutral loss ion types from backbone fragment ions and term them haphazardly. Therefore, eighteen different ion types are created (Table 3.2).

Table 3.2. Neutral loss ion types. On the right column, neutral ion types are generated from backbone fragment ions . (NH_3= 17.0307, H_2O= 18.0148).

x	-H_2O	=	**u**
y	- H_2O	=	**v**
z	- H_2O	=	**w**
x	-NH_3	=	**r**
y	-NH_3	=	**s**
z	-NH_3	=	**t**
a	- H_2O	=	**d**
b	- H_2O	=	**e**
c	- H_2O	=	**f**
a	-NH_3	=	**g**
b	-NH_3	=	**h**
c	-NH_3	=	**i**

For each m/z (ion) value from experimental tandem mass spectrum (the data set is indicated in section 2.1.1), in the study we constitute eighteen different "guesses" namely: "a, b, c, x, y, z, u (x-H_2O), v (y-H_2O), w (z-H_2O), r (x-NH_3), s (y-NH_3), t (z-NH_3), d (a-H_2O), e (b-H_2O), f (c-H_2O), g (a-NH_3), h (b-NH_3), i (c-NH_3)". Equating between the eighteen different guesses are used to find corresponding names for ions. It is possible that the ion can correspond to different names. For this reason, the most abundant ion type and its proportion are found. Then, the most abundant ion type is accepted as a label per m/z window. In this way, the **Named Spectrum** is generated.

By using the named ions (peaks) from the Named Spectrum, the named ions are translated into b-ion type by using fragment ion complementarity in order to create a b-ion only spectrum. It is called **b- ion Spectrum**. We expect that the abundance of b-ions could increase cumulatively because there may be multiple named ions which support one b-ion.

The main structure of the algorithm is illustrated in Figure 3.1 In order to explain it, three ions type which are a, b and c ions were created from MS-Product (2.1.7). As specified in Figure 3.1, named spectrum and b- ion spectrum are constituted. In b- ion spectrum, it is implied that a b- ion is found with high supports. Therefore, the algorithm, which is used, can account for missing b-ions in the experimental tandem mass spectrum and assign b-ions. Hence, b-ion ladder can be determined in order to assign peptide sequence of experimental tandem mass spectrum.

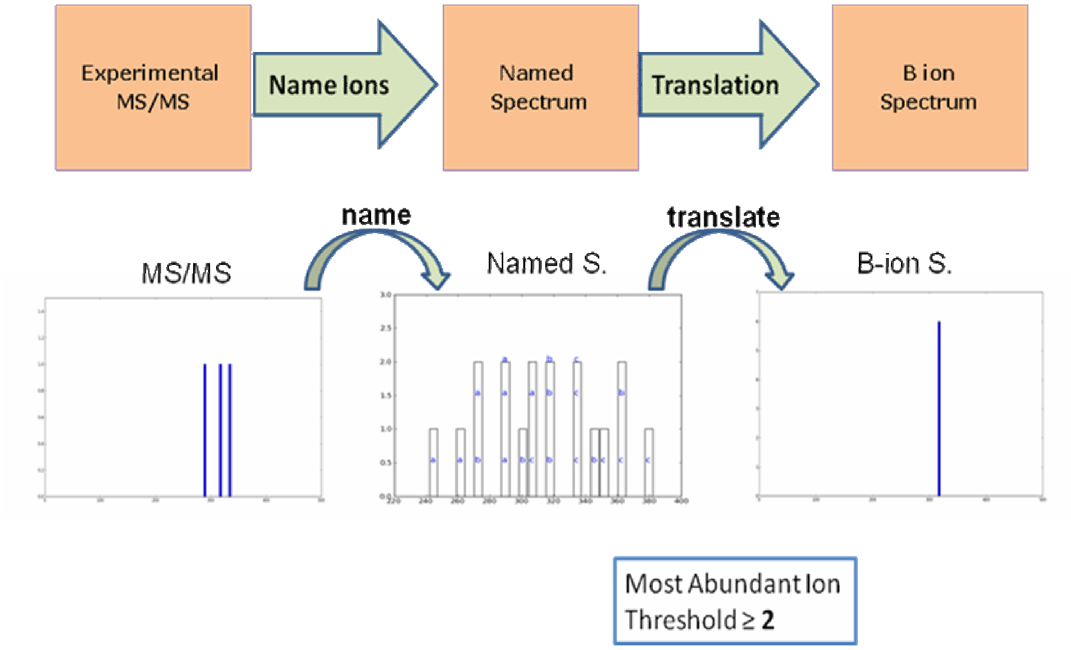


Figure 3.1. Ion Complementarity to generate Named and b-ion Spectrum. To reach b-ion ladder, firstly named spectrum is generated from MS/MS spectrum. As shown in the named s., some peaks are more interesting because they carry same ion type assignments. Then, these interesting peaks are translated into b-ion. Since they correspond same b-ion peak, the abundance of b-ion is increased.

Finding the full b-ion ladder with high supports and reducing the number of potential noise peaks as much as possible is a challenging task in the present study because of complexity of the experimental tandem mass spectrum and the named spectrum. Therefore, we need to add some steps to the named spectrum in order to decrease complexity and increase the supports of interesting peaks. Running sum and peak centering are applied to increase the supports of the important named peaks and lift them out from the potential noise. Then, filtering provides us to have the important named peaks with high supports and as few potential noise peaks as possible. After that, the named peaks are translated into b-ion peak by using fragment ion complementarity in order to generate b- ion spectrum. Because of the same reason with the named spectrum, b- ion spectrum requires peak centering and filtering. Then, the obtained b-ion peaks are traversed to the matrix application of Spectrum Graph. Ultimately, dynamic programming is used to find best peptide sequence from all possible amino acid sequence. The DENOVON algorithm is illustrated in Figure 3.2.

Naming Ions, Running Sum, Peak Centering, Filtering, Translation, Spectrum Graph and Dynamic Programming, which are steps of the DENOVON algorithm, are explained, in this section.

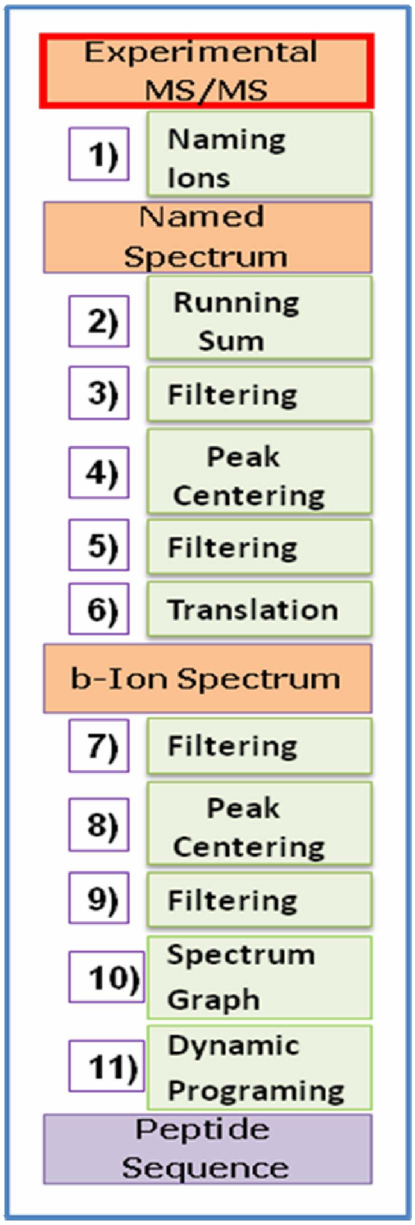


Figure 3.2. DENOVON algorithm. The steps of the DENOVON algorithm are shown to reach peptide sequence.

3.2. Naming Ions

The tandem mass spectra data set was measured by a LCQ MS/MS as indicated in section 2.1.1 As specified in section 2.2.1, we determine the best ion type set among eighteen different ion types for the data set. The best ion type set is **"b-y-h (b-ammonia)-c-i (c-ammonia)-u (x-water)-v (y-water)"** (As indicated in Table 3.2 and Table 3.1). Therefore, the best ion type set is used to generate the named spectrum from the experimental spectrum. In other words, as explained in section 3.1 and indicated in section 2.1.8, we assume one of the best ion types for each ion of the tandem mass spectrum then translate this ion type to the other best ion types by using fragment ion complementarity. In conclusion, the named spectrum is created for each spectrum.

In Figure 3.3, an example of a named spectrum is shown. In order to represent how running sum and peak centering are performed, the named spectrum was simplified. Therefore, the named spectrum which is indicated in Figure 3.3 is not real. In other words, this named spectrum was created manually because it is more comprehensible.

It is required to state that the actual figures which belong to the named spectrum and b spectrum are indicated in the results and discussion chapter.

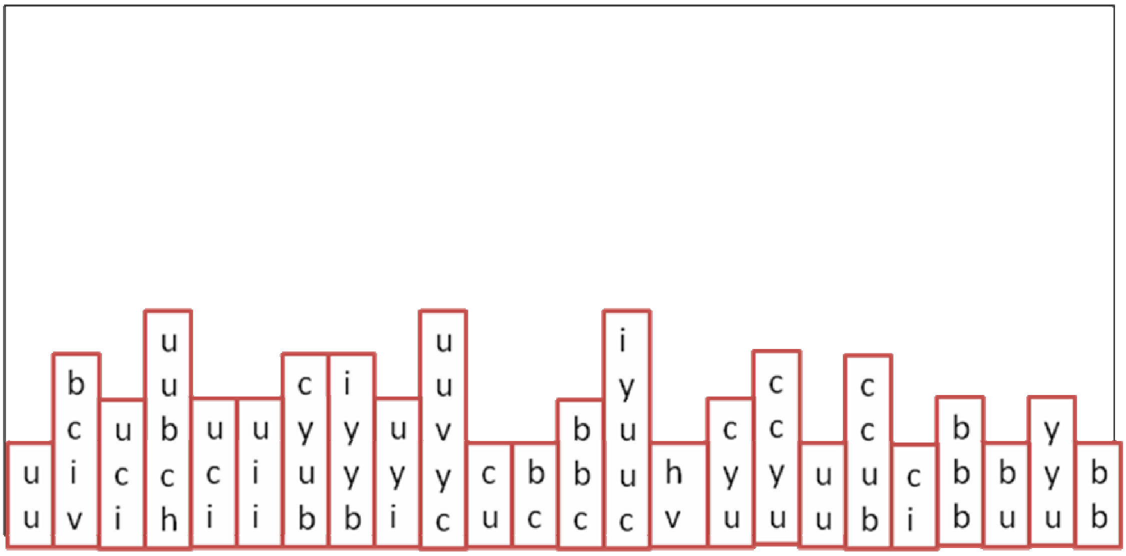


Figure 3.3. An Example of the Named Spectrum.

3.3. Running Sum

Firstly, in the present study, running sum is applied to the Named Spectrum in order to increase the abundance of interesting peaks and lift them out from the background of potential noise peaks.

Interesting peak means that it carries the same ion name type more than once. Therefore, the ion name assignment in such a peak is more valuable, because the same ion type provides support to name the peak. Moreover, these interesting peaks need to be considered.

Since we increase the abundance of ions in the experimental spectrum, the peaks in the named spectrum are adjacent to each other. Hence, we look for adjacent peaks to the interesting peaks whether they include the same ion type with the interesting peak or not.

Figures 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 3.10 and 3.11 show how running sum is performed in order to sharpen the interesting peaks. In a sliding window of size, the middle peak is considered if the most abundant name whether it is bigger than a threshold (Figure 3.6 and 3.10). If so, its adjacent peaks are checked whether they have same ion name assignment with the middle peak or not. After that, we add the value of the adjacent peak to the middle one if they have the same assignment (Figures 3.6, 3.7, 3.10 and 3.11). In this way, the support of interesting peaks is increased and the count of peaks in the named spectrum is decreased. Therefore, the complexity of the named spectrum is decreased (Figure 3.12).

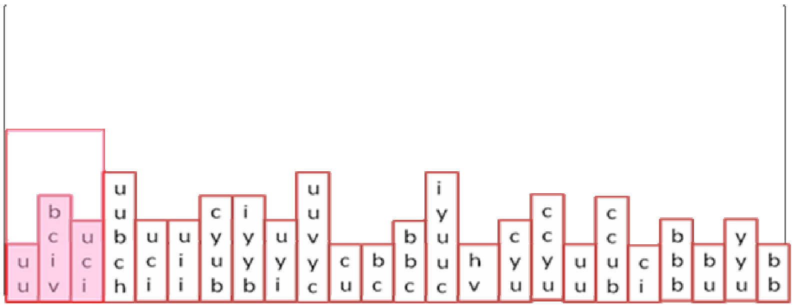


Figure 3.4. Step 1 of the Running Sum. Sliding window size is 0,7 and most abundant ion threshold is 2. Middle peak in the window size does not provide the threshold requirement.

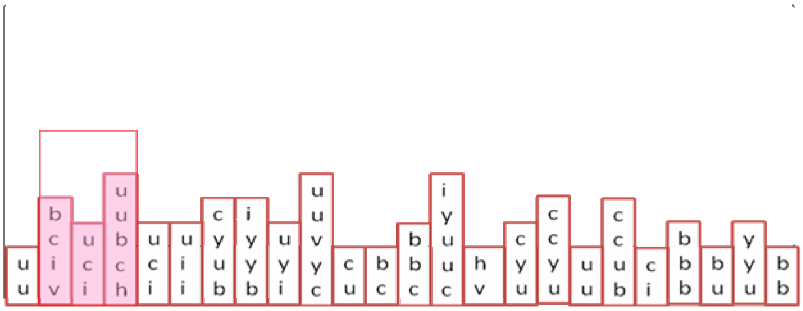


Figure 3.5. Step 2 of the Running Sum. Middle peak in the window size does not provide the threshold requirement.

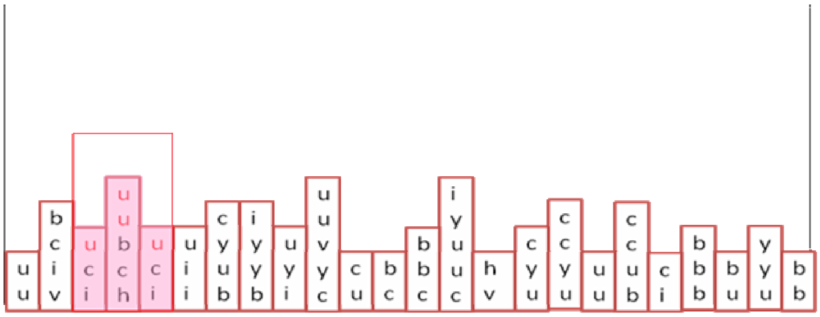


Figure 3.6. Step 3 of the Running Sum. Sliding. Middle peak in the window size provides the threshold requirement. Also, its adjacent peaks have same ion assignment which is ‘u’.

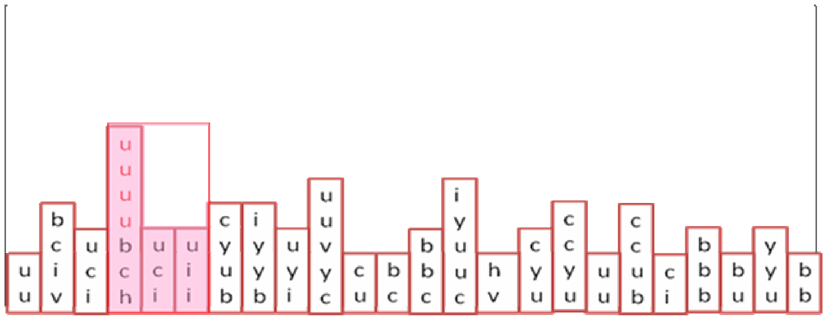


Figure 3.7. Step 4 of the Running Sum. In the former figure (Figure 3.6), a count of the abundance of interesting peak is two , after running sum it is increased four. In this sliding window, middle peak in the window size does not provide the threshold requirement.

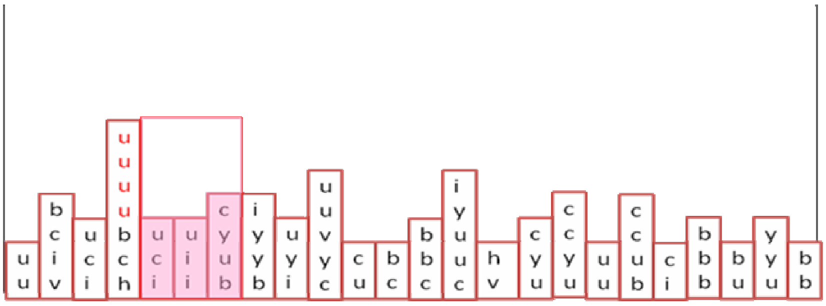


Figure 3.8. Step 5 of the Running Sum.

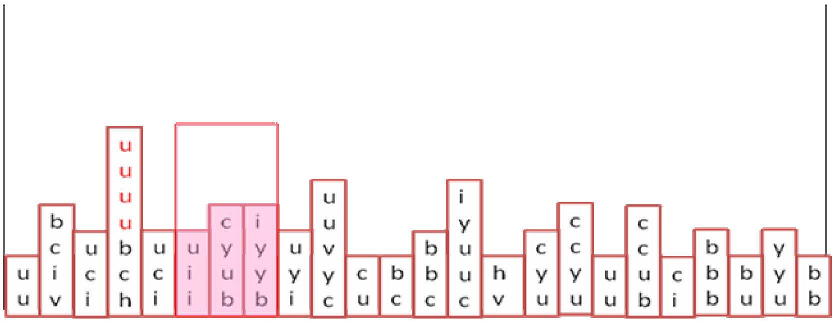


Figure 3.9. Step 6 of the Running Sum.

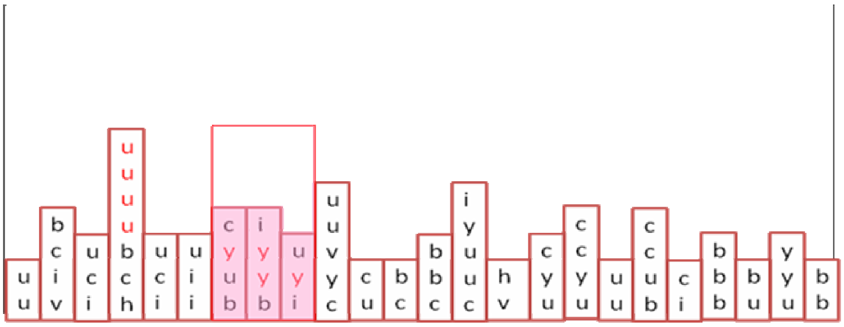


Figure 3.10. Step 7 of the Running Sum. Middle peak in the window size provides the threshold requirement. Also, its adjacent peaks have same ion assignment which is ‘y’.

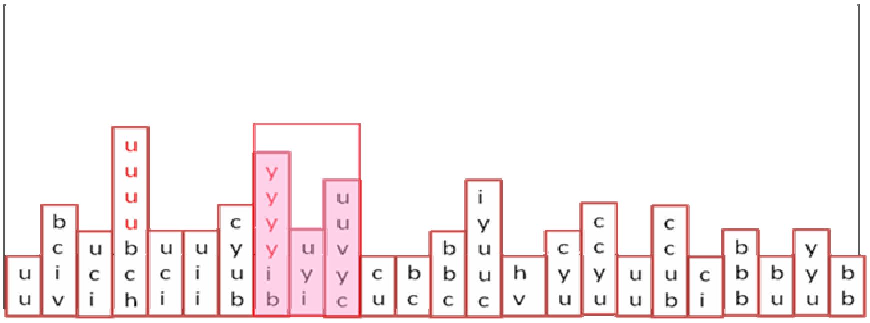


Figure 3.11. Step 8 of the Running Sum. In the former figure (Figure 3.10), a count of the abundance of interesting peak is two , after running sum it is increased four. In this sliding window, middle peak in the window size does not provide the threshold requirement. The running sum algorithm continues to the end of the named spectrum.

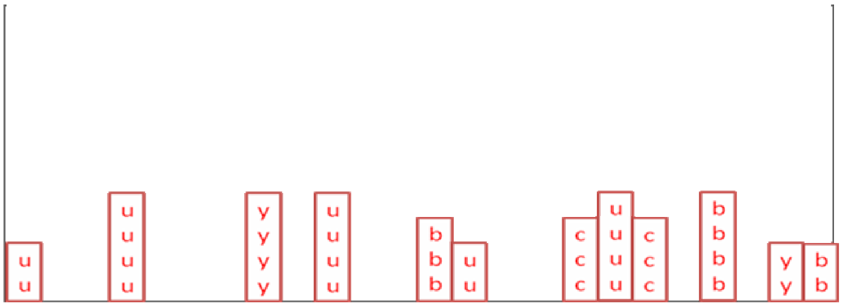


Figure 3.12. An Example of the Named Spectrum after applying the Running Sum. The complexity of named spectrum is decreased and the support of interesting peaks are increased.

Although, according to the DENOVON algorithm, the next step is filtering, peak centering is explained in order to use the same figures for clarification.

3.4. Peak Centering

After applying running sum, the count of peaks in the named spectrum is decreased. Moreover, as indicated in the Figure 3.12, the interesting peaks include only the same ion type.

The next step in the study is the peak centering to sharpen the interesting peaks and to remove the potential noise peaks.

In a sliding window, the adjacent peaks are checked whether they carry the same ion type or not. If the ion type is the same, the middle peak is calculated as a sum of m/z values. Then, the values of the adjacent peaks are added to the middle peak (Figures 3.12, 3.13, 3.15, 3.16 and 3.17).

As indicated in Figure 3.18, an implementation of the running sum and peak centering the named spectrum provides us to enhance the value of the interesting peaks. Therefore, the interesting peaks become more distinctive than the other potential noise peaks.

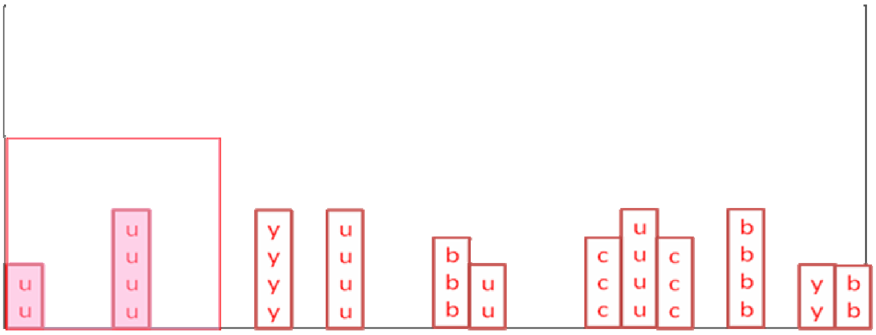


Figure 3.13. Step 1 of the Peak Centering. A sliding window size is 1.5 Da. In the window, the adjacent peaks carry same ion type. A middle peak is calculated and shown in Figure 3.14.

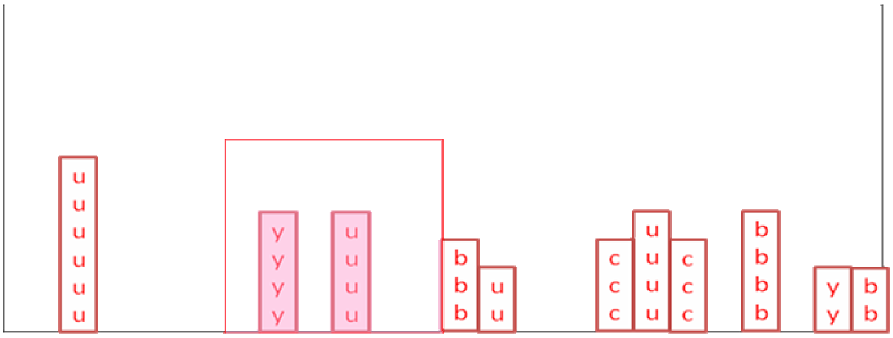


Figure 3.14. Step 2 of the Peak Centering. In the window, the adjacent peaks do not carry same ion type.

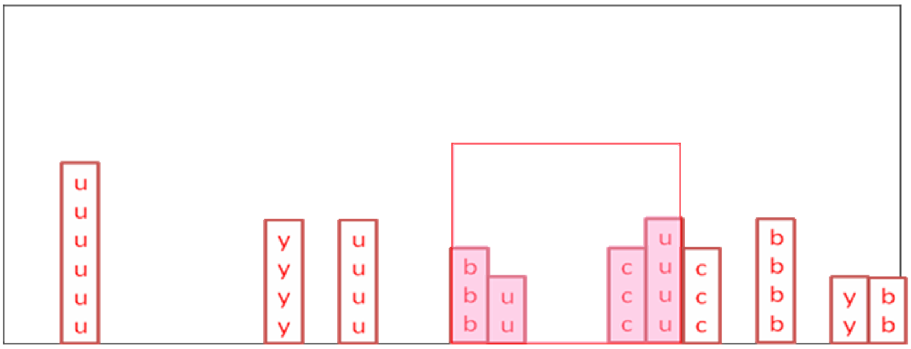


Figure 3.15. Step 3 of the Peak Centering. In the window, two of adjacent peaks carry same ion type. A middle peak is calculated and the values of the adjacent peaks are summed as shown in Figure 3.16.

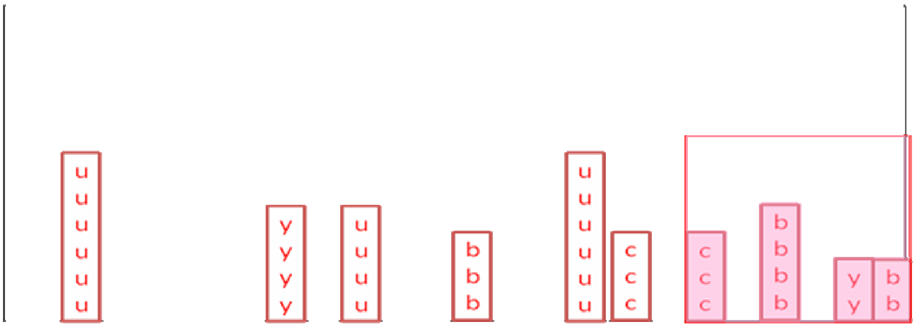


Figure 3.16. Step 4 of the Peak Centering. In the window, two of the adjacent peaks carry same ion type. A middle peak is calculated and the values of the adjacent peaks are summed as shown in Figure 3.17.

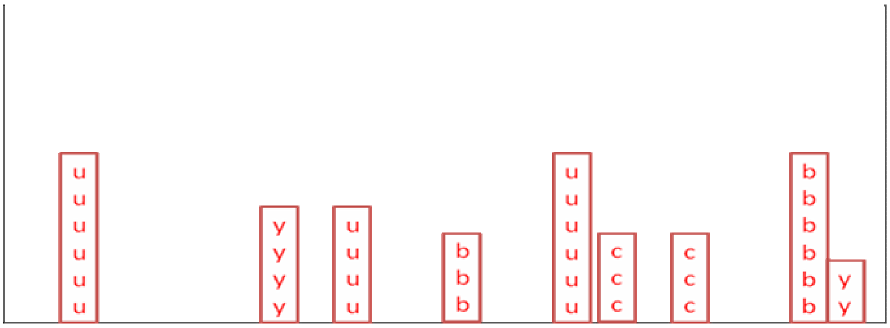


Figure 3.17. Step 5 of the Peak Centering. The most abundant ion threshold is three. Therefore, the peak which is smaller than the threshold is eliminated.

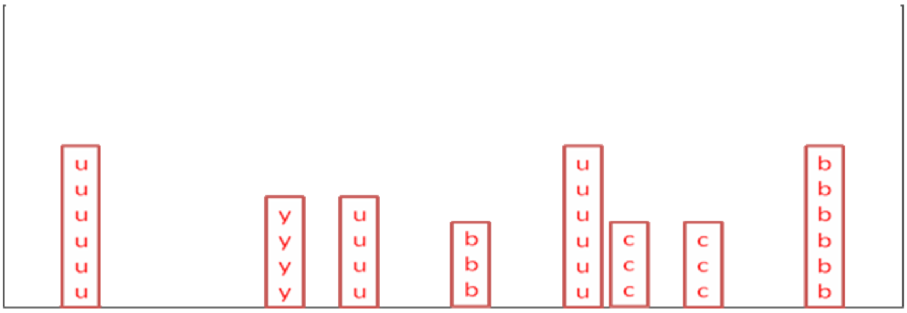


Figure 3.18. An Example of the Named Spectrum after applying the Running Sum and Peak Centering. The abundance of interesting peaks is increased and potential noise peaks are reduced.

As explained before in section 3.2 Naming Ions, the peak centering and the running sum were explained using the named spectrum. In addition, the named spectrum which is used in the figures was generated manually. Therefore, the actual versions of the named spectrum are shown in the Results and Discussion Chapter. In other words, the named spectrum, after applying the running sum the named spectrum and the peak centering the named spectrum are shown in the figures in the Results and Discussion Chapter.

In the present study, the next step is applied the named spectrum is filtering.

3.5. Filtering

Filtering is used in order to remove unexplained data from the spectrum in various *de novo* sequencing algorithms, for example, PEAKS which is indicated in section 1.1.3.1.

The named spectrum is more complicated than the figures which are shown in this chapter. Therefore, in addition to the running sum and the peak centering, filtering is required to obtain interesting peaks and separate them from potential noise peaks. We apply a sophisticated filtering approach to the named spectrum. To perform this, the named spectrum is divided into five windows. For each window, the specific threshold is determined statistically. In order to calculate the specific thresholds, all the data set was considered at the same time because the thresholds which belong to five windows are then used to cut to the named peaks. If support of the named peak in the window is bigger than the threshold, the named peak is taken because of its importance. If not, the named peak is removed.

In the present study, the next step of the algorithm is the translation. It translates the named peaks into the b-ion Spectrum.

3.6. Translation, Peak Centering and Filtering

As shown in the Figure 3.2 and as indicated in 3.1, the translation is a latter step in the algorithm. The named peaks in the named spectrum which are bigger than a threshold are translated into the b- ion peaks by using fragment ion complementarity in order to generate a spectrum which includes only b-ions.

The b-ion spectrum should include a full b-ion ladder and as few as possible potential noise peaks. Since it is necessary to decrease the count of the potential noise peaks in the b- ion spectrum, peak centering and filtering are applied to the b- ion spectrum.

As explained in the section 3.4, the same logic can be used for the b- ion spectrum because the aim of application of the peak centering is the same for both the named spectrum and the b-ion spectrum. Unlike the named spectrum, the b-ion spectrum includes only b-ion peaks. Therefore, in order to find the middle peak, the support of the b-ion is considered. Then, the support of the important peak is increased.

Afterwards, the filtering approach is conducted on the b-ion spectrum in order to obtain full b-ion ladder and to reduce the potential noise peaks as much as possible. In the section 3.5, the filtering is explained elaborately. Also, the reason for implementation of the filtering is the same for both the named spectrum and the b-ion spectrum.

3.7. Spectrum Graph and Dynamic Programming

An ideal tandem mass spectrum does not have unexplained peaks and contains only b-ions and y-ions, and every ion peak has similar abundance. Eventually, the filtered b-ion spectrum almost reaches to the ideal tandem mass spectrum because it includes b-ion ladder and potential noise b-ion peaks as few as possible. Therefore, they can be used to find the peptide sequence.

As explained in section 1.1.2.2, there are two major steps in order to find the peptide sequence. The first step is to generate a pool of candidate sequences. The second step is to design a scoring function to select the best candidate from the pool. To do these, most *de novo* algorithms create a spectrum graph and then apply a dynamic programming algorithm to find paths in the graph that represent peptide candidates (Alves & Yu, 2005), (Chen, Kao, Tepel, Rush, & Church, 2001), (Dancik, Addona, Clauser, Vath, & Pevzner, 1999), (Lu & Chen, 2003) and (Ma, Doherty-Kirby, & Lajoie, 2003).

In the present study, the b-ions and the precursor mass of the tandem mass spectrum are traversed to the matrix application of spectrum graph. Each peak in the b-ion spectrum serves as vertex. A directed edge connects two vertices in the spectrum graph if the difference between the vertices' masses is the mass of a single amino acid. Precision (closeness of fit for amino acid) and naming confidence of vertices make up the weights for the edges and vertices. This, then, enables us to determine the highest scoring path and offer it as the amino acid sequence that gave rise to the MS/MS spectrum.

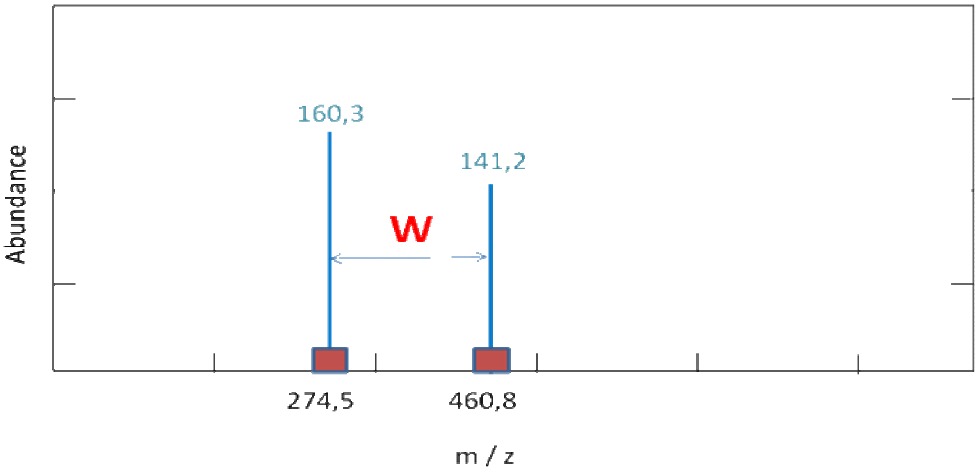


Figure 3.19. Example peaks of b-ion Spectrum. 🟥 shows the absolute error. W= monoisotopic mass 186.07931, W=average mass 186.2132.

In Figure 3.19, two example b-ion peaks of a spectrum are shown to explain how the scoring function is calculated. According to the figure, the m/z value of the first peak is 274,5 (V_1) and its naming support is 160,3. Similarly, 460,8 (V_2) is the m/z value of the second peak and 141,2 is its naming support. The first parameter for calculating scoring function is an absolute error. As explained above, the mass difference between two vertices is 186,3. Although the mass of W (Trp) is not equal to the difference exactly, the edge (E) connects two vertices because an absolute error is considered as 0,7 Da. The second parameter is the naming confidence of the vertices. To do that, the average of the naming confidences is calculated. Therefore, these parameters create a weight to the edge (Figure 3.20). The scoring provides us to determine the highest scoring path among all possible peptide sequences and offer it as the amino acid sequence of the MS/MS spectrum.

3.8. Class Diagram

A class diagram is a type of static structure diagram that describes the structure of a system by showing the system's classes and the relationships between the classes.

Figure 3.20 shows the most important classes which are generated for the present study and the relationships between these classes.

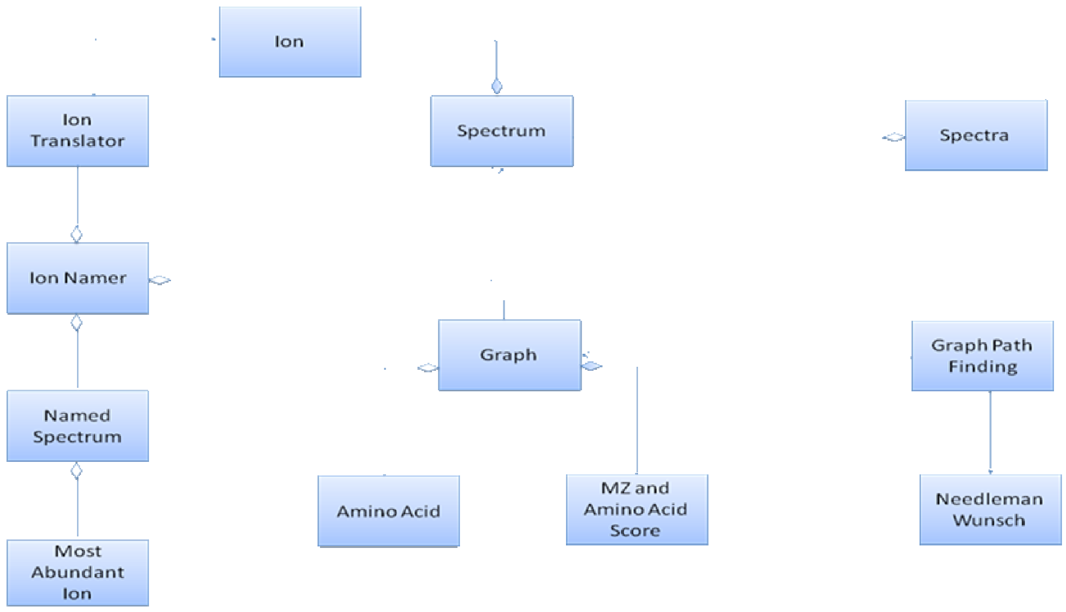


Figure 3.20. Class Diagram.

CHAPTER 4

RESULTS AND DISCUSSION

In Chapter 3, the DENOVON algorithm is explained in order to obtain b-ion ladder and as few noise peaks as possible from an experimental tandem mass spectrum. This simplifies the complexity of experimental spectrum and b-ions which are important for peptide identification can be found.

4.1. Named Spectrum

The Figure 4.1 is from the published data set (Keller et al., 2002) in section 2.1.1 As shown in the Figure 4.1, the number of peaks is 221.

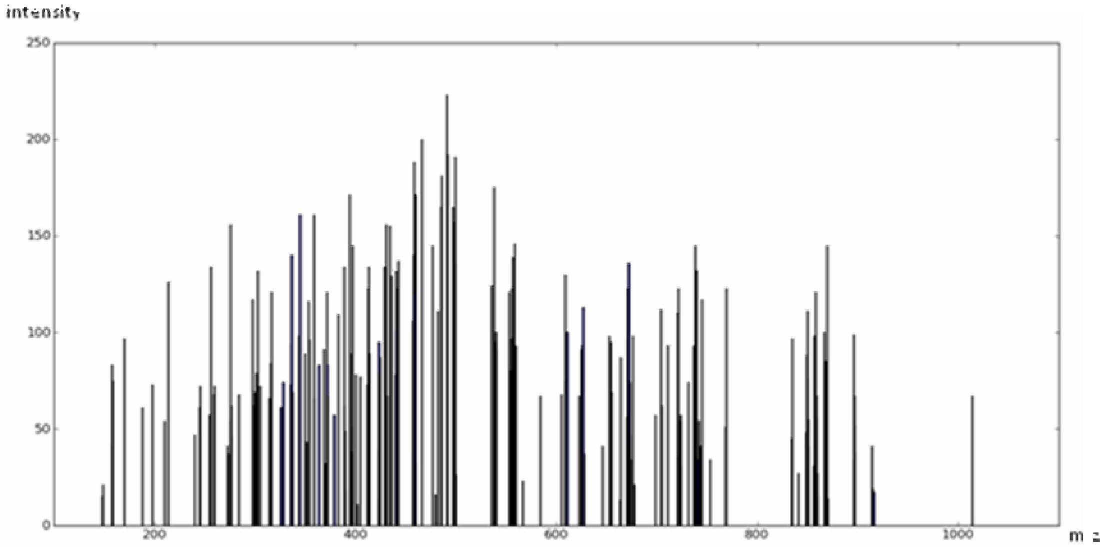


Figure 4.1. Experimental Tandem Mass Spectrum. It belongs to the data set as indicated in section 2.1.1 is an experimental tandem mass spectrum. Before processing the algorithm, peak count of the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum is 221.

As can be seen in Figure 4.2, by using ion complementarity the named spectrum was generated from the experimental tandem mass spectrum as explained in section 3.2 Since the number of potential noise peaks seems high and interesting peaks needed to be increased in abundance, running sum, filtering and peak centering were applied to the named spectrum.

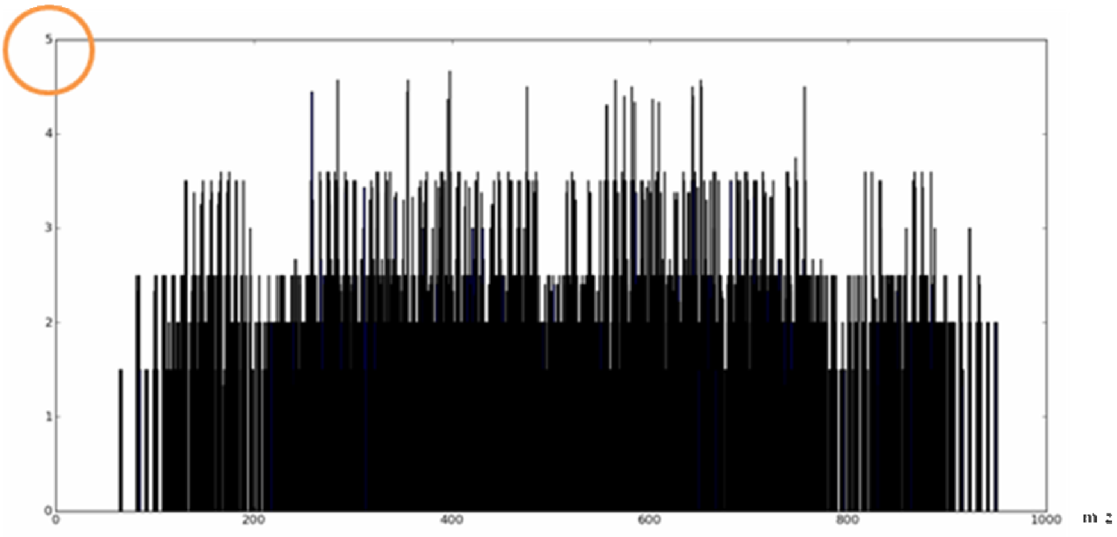


Figure 4.2. Named Spectrum. A proportion of the most abundant ion type is referred as y-axis. The proportion shows naming confidence for each peak. After naming ions, peak count increases from 221 to 4635 for the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum.

4.2. Running Sum the Named Spectrum

After applying running sum (section 3.3) to the named spectrum (Figure 4.2), the support of interesting peaks is increased and the number of potential noise peaks is decreased as illustrated in Figure 4.3 To demonstrate the change of the proportion of most abundant ion type of interesting peaks is shown as circle in the Figures 4.2, 4.3, 4.4 and 4.5. In addition, in order to point out the change in number of peaks, the Peak Count is also illustrated in all figures in this chapter.

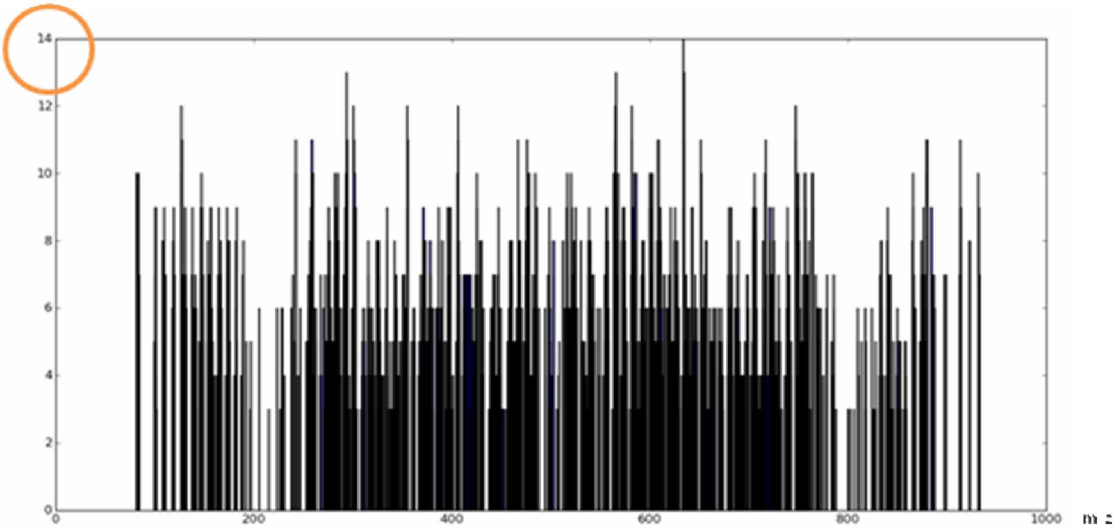


Figure 4.3. After running sum, appearance of the Named Spectrum. Peak count reduces from 4635 to1423 for the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum. Also, the proportion of the most abundant ion type increases from 5 to 14.

4.3. Filtering the Named Spectrum

As explained in section 3.5, according to specific thresholds for each window, the number of peaks was reduced as shown in the Figure 4.4. Since filtering provides a decrease the count of possible noise, a change in proportion of the most abundant ion type does not occur.

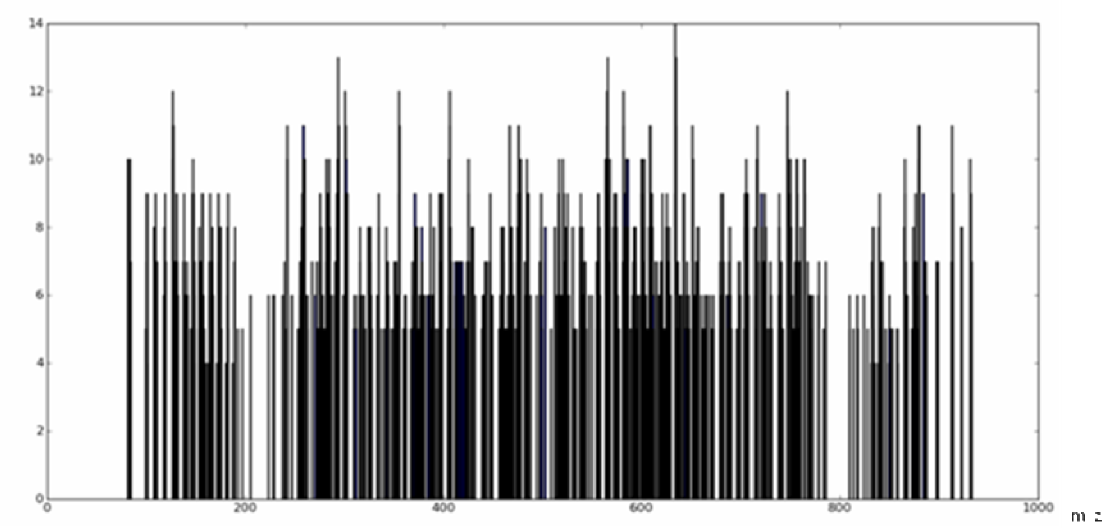


Figure 4.4. After filtering, appearance of the Named Spectrum. Peak count decreases from 1423 to 1036 for the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum.

4.4. Peak Centering the Named Spectrum

As explained in the section 3.4, the support of interesting named peaks is increased in order to lift them out from the potential noise peaks (Figure 4.5).

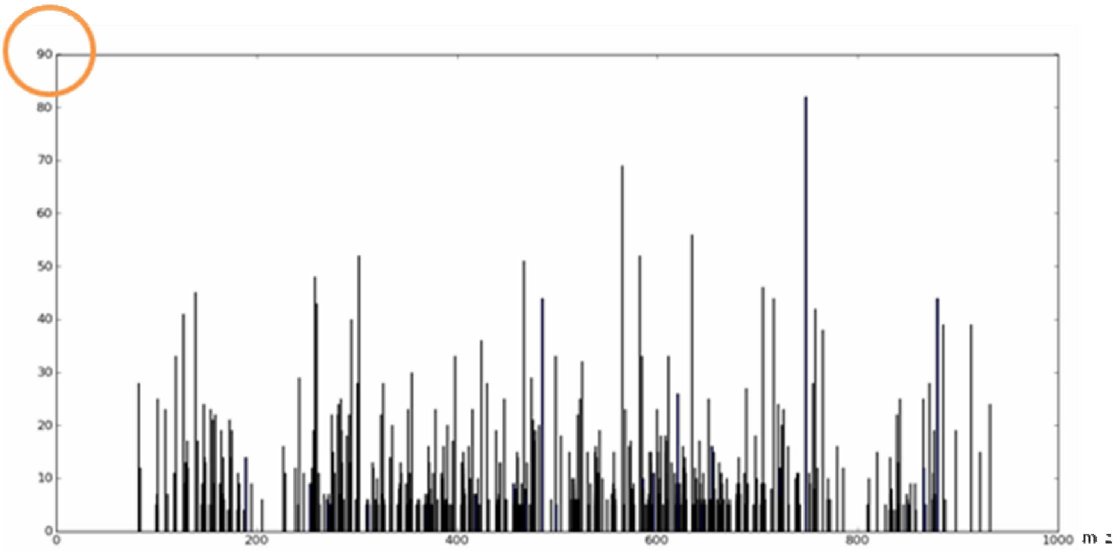


Figure 4.5. After peak centering, appearance of the Named Spectrum. Peak count decreases from 1036 to 642 for the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum. In addition, naming confidence increases from 14 to 90.

4.5. Filtering the Named Spectrum

The next step in the present study is also filtering as indicated in Figure 3.2. As shown in Figure 4.6, the complexity of the named spectrum was reduced.

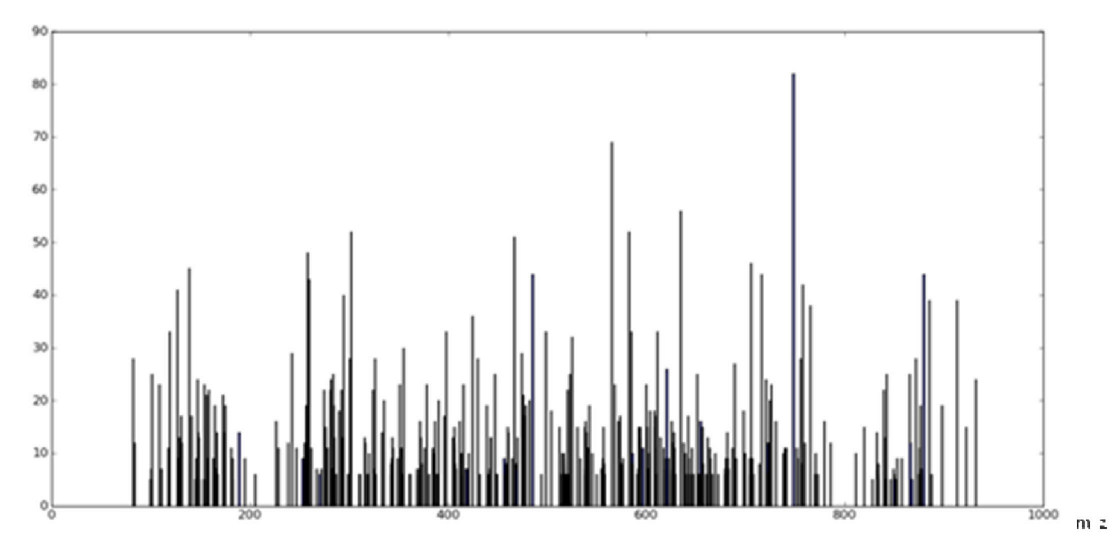


Figure 4.6. After filtering, appearance of the Named Spectrum. Peak count decreases from 642 to 456 for the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum.

4.6. b-ion Spectrum

By exploiting fragment ion complementarity, the named peaks are, then, translated into b-ion type in order to generate the spectrum which includes only b-ions as explained in section 3.1 and section 3.6.

Cumulative b-ion abundance is given on the y-axis in b-ion Spectrum. We expected that the abundance of b-ions could increase and peak count could decrease because there may be multiple named ions which support one b-ion. Therefore, as shown in Figure 4.7, peak count is less than in Figure 4.6 and the abundance of y-axis is higher than in Figure 4.6 Both results support our expectation.

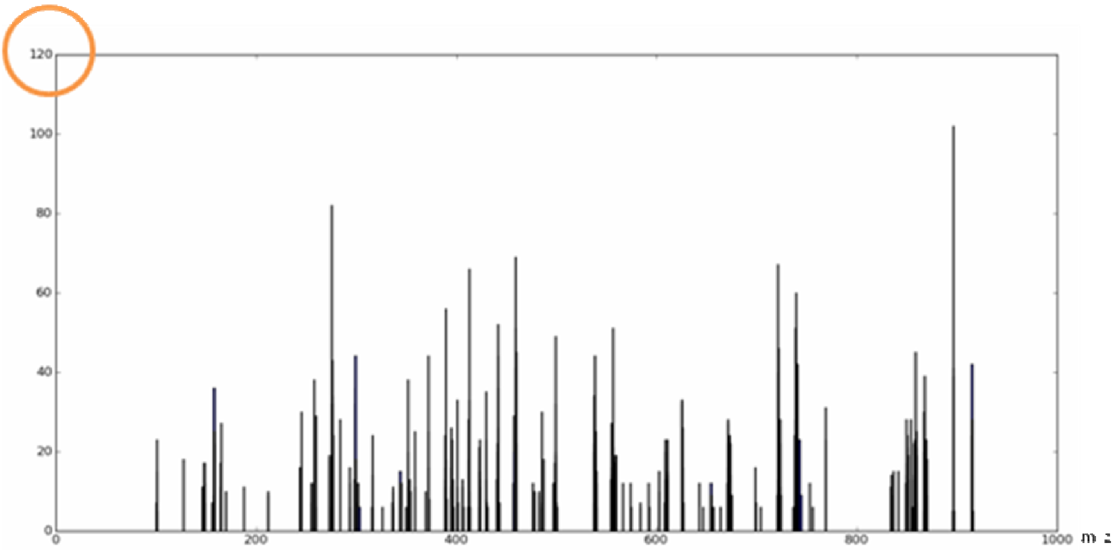


Figure 4.7. b-ion Spectrum. The b-ion abundance increases from 90 to 120. In addition, peak count reduces from 456 to 303 for the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum.

4.7. Filtering the b-ion Spectrum

Figure 4.8 shows elimination of the potential noise peaks according to determined threshold for each window.

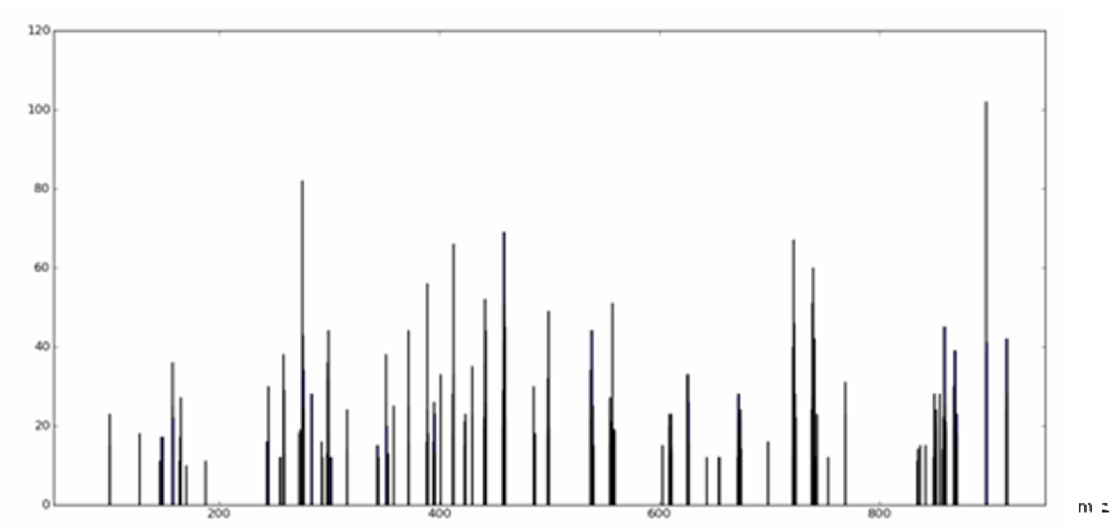


Figure 4.8. After filtering, appearance of the b-ion Spectrum. Since peak count decreases from 303 to 189, the complexity of VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum reduces.

4.8. Peak Centering the b-ion Spectrum

As illustrated in the Figure 4.9, the support of interesting b-ions is increased and the number of potential noise peaks is reduced.

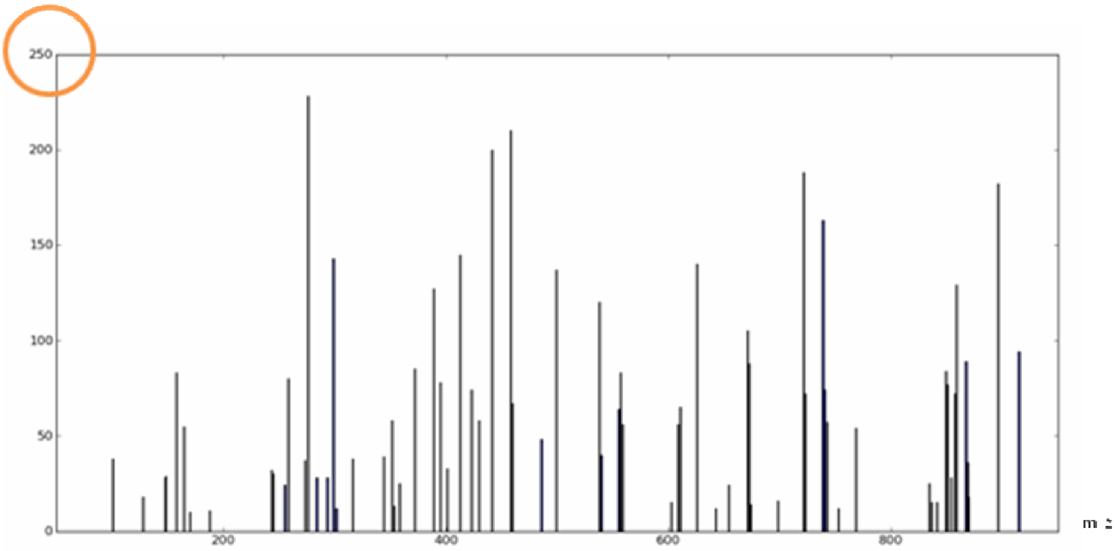


Figure 4.9. After peak centering, appearance of the b-ion Spectrum. The b-ion abundance increases from 120 to 250. In addition, peak count reduces from 189 to 70 for the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum.

4.9. Filtering the b-ion Spectrum

As shown in Figure 3.2, the next step is filtering the b-ion spectrum. The b-ion peaks which are smaller than determined threshold were removed from the b-ion Spectrum (Figure 4.10).

We achieved to gain a b-ion ladder which enables us to assign a peptide sequence to an experimental tandem mass spectrum. In addition, in contrast to an experimental tandem mass spectrum, peak count is reduced and b-ions are enriched as shown in Figure 4.10 Therefore, the unexplained peaks from experimental tandem mass spectrum are decreased.

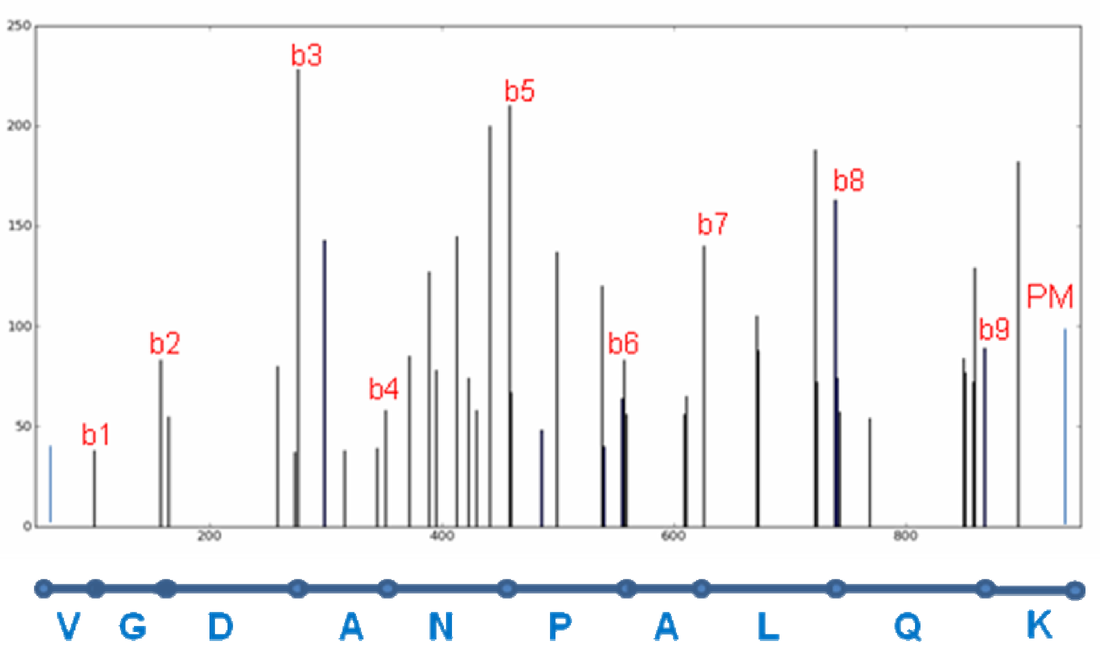


Figure 4.10. After filtering, appearance of the b-ion Spectrum. Peak count reduces from 189 to 70 for the VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum. In addition, b-ion ladder which provides to find peptide sequence is shown in the figure. PM is a precursor mass of the spectrum.

While we managed to gain full b- ion ladder for 10 spectra which were assigned with short peptide sequence, we could not find full b- ion ladder for 740 spectra which was designated with long peptide sequence. Since LCQ tandem mass spectra have a low resolution for low m/z range and for high m/z range, obtaining the first b-ion and the last b-ion is a challenging task. As shown in Figure 4.10, the last b-ion which is b_{10} for

VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta tandem mass spectrum is not found. However, nonexistence of the last b-ion does not cause a problem to identify peptide because it would be just a precursor mass as indicated in Figure 4.10. The results which belong to VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta are shown in Table 4.1. and Table 4.2.

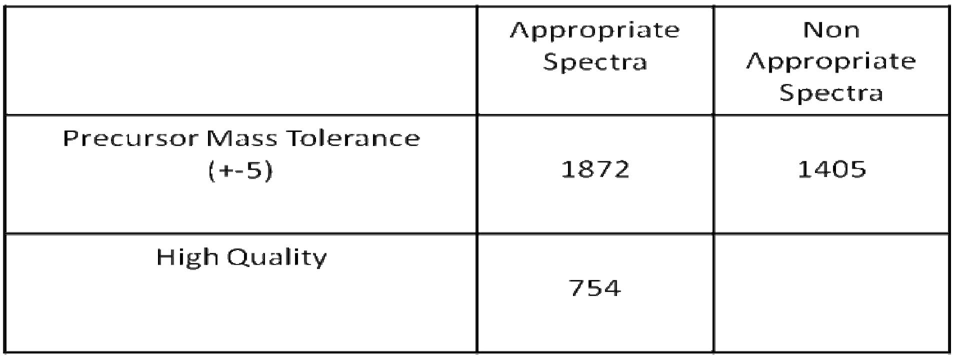
Table 4.1. After Processing DENOVON. After processing the algorithm, peak count is reduced 5 times in VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum. In addition, while found b-ions/ theoretical b-ions ratio is 0,7 in the experimental tandem mass spectrum, after processing the algorithm the ratio is increased to 0,9. It means that some missing b-ions in the experimental tandem mass spectrum are found.

	Peak Count	Found b-ions/Theoretical b-ions
Before Processing	221	0,7
After Processing	43	0,9

Table 4.2. Found and Expected Sequence. The relative distance is 6 between expected and found peptide sequence in VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta spectrum.

	Expected	Found	Relative Distance
Sequence	VGDANPALQK	VGDAPGALQY	6

Table 4.3. Evaluation quality of LCQ Experimental Tandem Mass Spectra. The spectra belong to published data set (Source: Keller et al, 2002). According to precursor mass tolerance, 1405 spectra were eliminated from 3000 spectra. Then, the high quality spectra were selected from 1872 appropriate spectra in order to evaluate the DENOVON.



	Appropriate Spectra	Non Appropriate Spectra
Precursor Mass Tolerance (+-5)	1872	1405
High Quality	754	

Table 4.4. Results of high quality LCQ Experimental Tandem Mass Spectra. The spectra belong to published data set (Source: Keller et al, 2002). After processing the algorithm, the peak count of 754 spectra is reduced 2.5 times. In addition, before and after processing of DENOVON on 754 spectra, found b-ions/ theoretical b-ions ratio is similar. All mean that while processing of the spectra we do not lose b-ion ladder which is important for peptide identification. In addition, the complexity of the spectra is reduced.

	Appropriate and High Quality Spectra	
	Peak Count	Found b-ions/Theoretical b-ions
Before Processing	210,2	0,70
After Processing	81,5	0,69

As indicated in Table 4.3, DENOVON algorithm was evaluated on 754 spectra which belong to the published data set. As shown in Table 4.4, 0,69 is the ratio of found b-ions/theoretical b-ions for the 754 spectra. The ratio indicates that we reached b-ion ladder after processing our algorithm. In addition, the peak count of the spectra was reduced significantly. Therefore, then, we can find amino acid sequence of peptide of these spectra. Moreover, the DENOVON algorithm established missing b- ions for 21 spectra of the 754 spectra. Hence, not only correct name assignment to ion and reducing the complexity of experimental tandem mass spectrum was managed, but also some missing b- ions were found for some spectra. It can be indicated for the spectrum "VGDANPALQK_sergei_digest_A_full_01.0485.0487.2.dta" because the ratio of found b-ions/theoretical b-ions is increased after processing the algorithm as shown in Table 4.1.

In order to evaluate our DENOVON algorithm, it is necessary to compare to another algorithm on the same data set. Therefore, we compared our algorithm to PepNovo algorithm (in section 1.1.3.2) which is one of the popular *de novo* sequencing algorithm. We used the same 754 experimental tandem mass spectra for evaluation of two different algorithms. However, PepNovo found peptide sequences for only 72 spectra of 754 spectra. Hence, by using PepNovo algorithm, we could not reach any result for 682 spectra of the 754 spectra. Differently, DENOVON algorithm managed to obtain peptide sequence for all of the 754 spectra. In order to compare our algorithm to PepNovo, we needed to divide the spectra in two as 72 spectra and the rest of the spectra due to limitation of PepNovo. In addition, since PepNovo can find peptide sequence for only limited number of the spectra, for the rest of the spectra we used b-ion spectra which are generated by DENOVON algorithm form experimental spectra. In Figure 4.11, results of PepNovo algorithm are shown for 72 spectra of the experimental spectra, 72 spectra of b-ion spectra, the rest of the b-ion spectra. Relative distance indicates accuracy of the algorithm for the spectra. If relative distance is high for the spectra, it means the accuracy of peptide assignments is low. The correctness of DENOVON algorithm for the experimental spectra is shown in Figure 4.12.

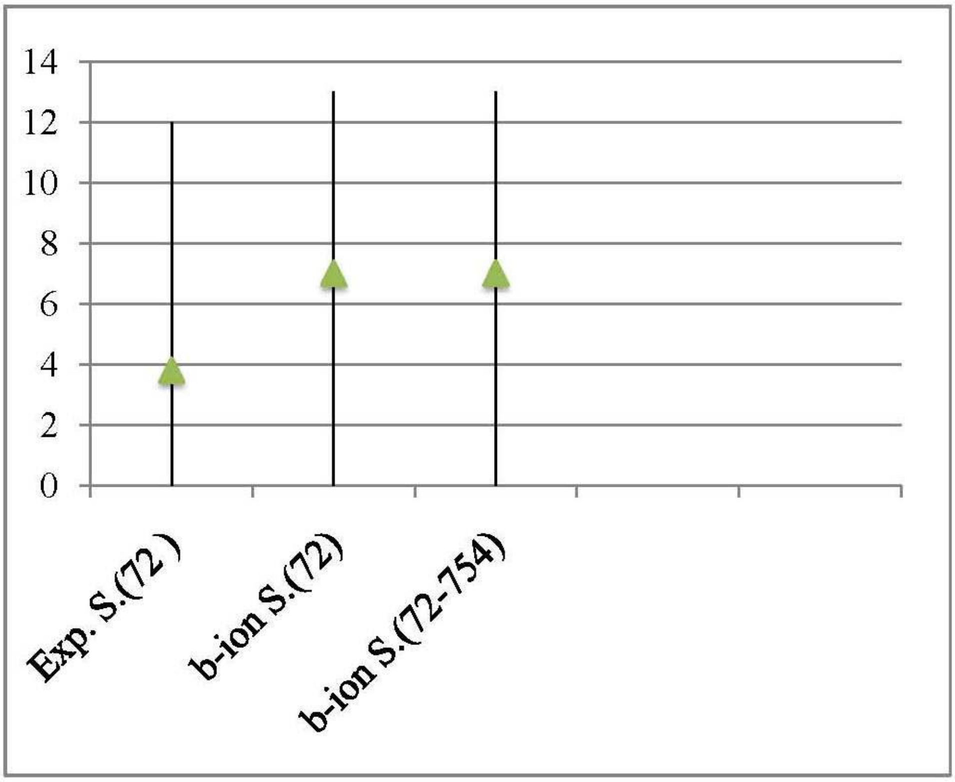


Figure 4.11. Relative Distance of the PepNovo for LCQ MS/MS Spectra which are published data set (Source: Keller et al, 2002). ▲ shows average relative distance between expected and found peptide sequences. Exp.S. (72) has average relative distance as 3,8, minimum relative distance as 0 and maximum relative distance as 11. b-ion S. (72) has average relative distance as 7, minimum relative distance as 0 and maximum relative distance as 13. b-ion S. (72-754) has average relative distance as 7, minimum relative distance as 0 and maximum relative distance as 13.

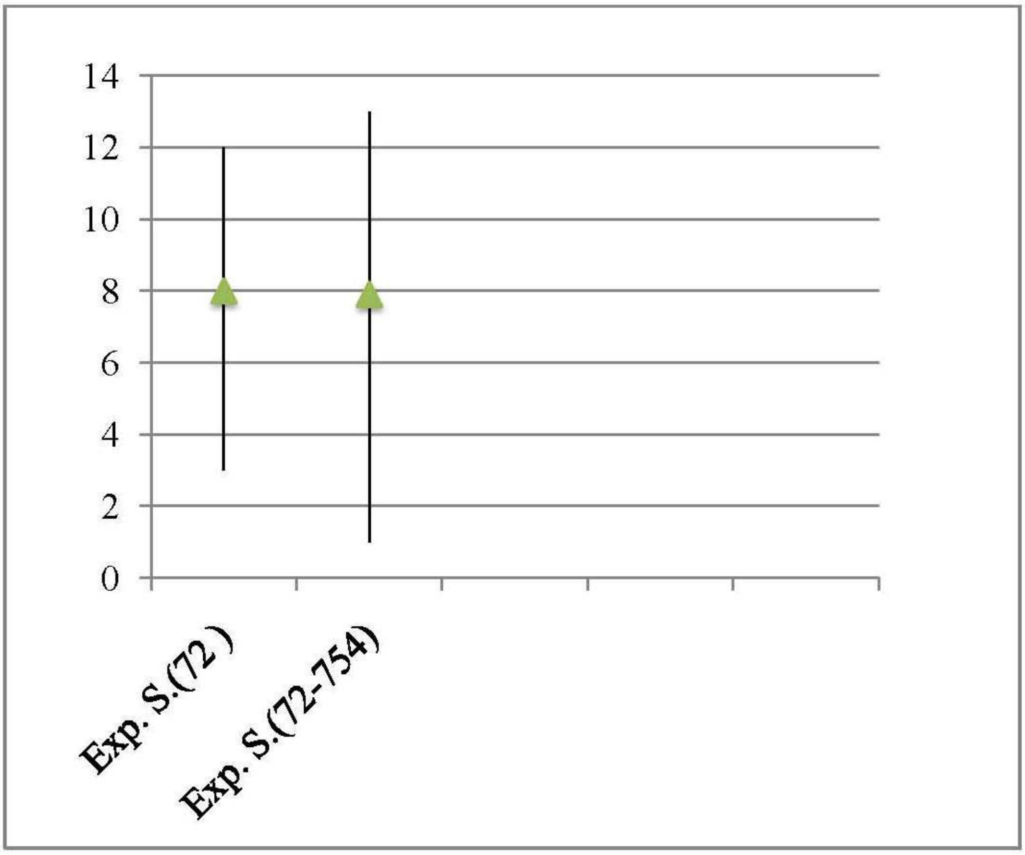


Figure 4.12. Relative Distance of the DENOVON for LCQ MS/MS Spectra which are published data set (Source: Keller et al, 2002). ▲ shows average relative distance between expected and found peptide sequences. Exp.S. (72) has average relative distance as 8, minimum relative distance as 3 and maximum relative distance as 12. Exp S. (72-754) has average relative distance as 7,9, minimum relative distance as 1 and maximum relative distance as 13.

As indicated in Figure 4.11, although the accuracy of the PepNovo seems good for the 72 spectra of the experimental tandem mass spectra, it cannot generate any peptide assignment to the rest of the experimental tandem mass spectra. Therefore, we managed to evaluate the accuracy of PepNovo by using b-ion spectra which are generated by the DENOVON. In Figure 4.12., DENOVON algorithm is evaluated on the experimental tandem mass spectra. Both 72 spectra and the rest of the spectra have similar correctness to peptide assignment. Also, they have similar relative distance with b-ion spectra in Figure 4.11 The results show that the DENOVON can find peptide sequence to experimental tandem mass spectra. In addition, b-ion spectra can be used to reach peptide assignment for the PepNovo algorithm. In future, the accuracy of the DENOVON will be increased to reach full sequence of peptides.

CHAPTER 5

CONCLUSION

Mass spectrometry based proteomics is a remarkably powerful approach in to identify proteins in complex biological samples. Since correct peptide assignment is a key step for protein identification, huge number tandem mass spectra are frequently generated in proteomics experiments and they require challenging data analysis. Despite recent improvements in identification methods, a significant number of spectra remain unidentified due to post translational modifications, mutations and incompleteness of protein databases. These spectra may represent meaningful biological information and are potentially identifiable with *de novo* sequencing because it can assign amino acid sequence of peptide to tandem mass spectrum without the need of a protein database.

Unknown fragment ion types, unknown charges, incomplete fragmentation, unexplained data, isotopic ions, and machine errors are limiting algorithms while analyzing. Therefore, up to this point, different *de novo* sequencing algorithms were developed to manage some difficulties while identifying peptides.

The aim of this study was to enhance the accuracy of *de novo* sequencing algorithms. Naming of fragment ions is one common step in all *de novo* sequencing algorithms because it is essential to know which peak represents which ion type in order to traverse a spectrum graph to find an amino acid sequence of peptide. Different approaches have achieved to name b-type ions and y-type ions.

We have presented a new approach which provides the naming of not only b-type ions and y-type ions but also other arbitrary ion types. By exploiting fragment ion complementarity, the detection of full b-ion ladders for tandem mass spectra data set was achieved in some cases. In addition, unexplained data in the data set was reduced substantially. Moreover, missing fragments in some of the data set have been determined by using other named ion types. In order to evaluate the accuracy of DENOVON algorithm, our algorithm was compared to PepNovo algorithm. While PepNovo can find peptide assignments to 72 experimental spectra of the data set, the DENOVON algorithm can assign peptide sequences to 754 experimental spectra of the data set. In order to reach full sequence of peptides, the correctness of DENOVON will be enhanced in future.

It is possible that the development of this novel method, and the enhancement of the remaining approach, to *de novo* sequencing of tandem mass spectra will have great influence on mass spectrometry and proteomics since owing to properly named fragment ions more difficult task like the *de novo* sequencing of tandem mass spectra with post translational modifications can be enabled. This may impact clinical research by enabling the direct analysis of proteins without prior knowledge of their origin. It is anticipated that DENOVON algorithm might popularize in proteomics laboratories because of overcoming the limitation of database search.

REFERENCES

A. M. Falick, H. W., Medzihradszky, K. F., Baldwin, M. A., & Gibson, B. W. (1993). Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry , 4* (11), 882-893.

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

Agilent Spectrum Mill MS Proteomics Workbench. (2010). Retrieved April 16, 2011, from Spectrummill web site: http://spectrummill.mit.edu/millhtml/sm_instruct/utilityman.htm#useMSProd

Allmer, J. (2011). *De Novo* Sequencing Algorithms for LC-MS/MS Spectra.

Allmer, J. (2006). Development of algorithms for peptide identification from mass spectrometric data in genomic databases. *Ph.D.thesis* .

Alves, G., & Yu, Y. (2005). Robust accurate identification of peptides (RAId): deciphering MS2 data using a structured library search with *de novo* based statistics. *Bioinformatics , 21* (19), 3726-3732.

Amino acid masses. (2006). Retrieved May 12, 2010, from http://education.expasy.org/student_projects/isotopident/htdocs/aa-list.html

Bandeira, N. (2007). Spectral networks: a new approach to *de novo* discovery of protein sequences and posttranslational modifications. *BioTechniques , 42*, 687–695.

Bartels, C. (1990). Fast algorithm for peptide sequencing by mass spectrometry. *Biomed. Environ. Mass Spectrom , 19*, 363-368.

Bern, M., & Goldberg, D. (2005). EigenMS: *de novo* analysis of peptide tandem mass spectra by spectral graph paritioning. *RECOMB , 3500*, 357-372.

Bern, M., Cai, Y., & Goldberg, D. (2007). Lookup Peaks: A Hybrid of *de Novo* Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. *Anal. Chem , 79*, 1393-1400.

Bilsel, T. (2009, January 30). *tayfunbilsel*. Retrieved April 11, 2011, from tayfunbilsel web site: http://www.tayfunbilsel.com/?page_id=18

Burlingame, A. (n.d.). *Protein Prospector*. Retrieved April 16, 2011, from Protein Prospector: http://prospector.ucsf.edu/prospector/mshome.htm

Chait, B. (2006). Mass Spectrometry: Bottom-Up or Top-Down? *Science , 314*, 65-66.

Chan, P., & Lee, R. (1996). *The Java Class Libraries: An Annotated Reference* (Vol. 1). Boston, MA, USA: Addison-Wesley Longman Publishing Co.

Chen, T., Kao, M., Tepel, M., Rush, J., & Church, G. (2001). A dynamic programing approach for *de novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology , 8*, 325-337.

Chen, T., Kao, M., Tepel, M., Rush, R., & Church, G. M. (2001). *A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry.* (Y. Huang, & H. Wang, Performers) Taiwan.

Daconta, M. C. (1996). *Java for C/C++ Programmers* (Vol. 1). New York: John Wiley & Sons.

Dancik, V., Addona, T., Clauser, K., Vath, J., & Pevzner, P. (1999). *De novo* peptide sequencing via tandem mass spectrometry. *J. Comp. Biol , 6*, 327-342.

Dancik, V., Addona, T., Clauser, K., Vath, J., & Pevzner, P. (1999). *De novo* peptide sequencing via tandem mass spectrometry. *J Comput Biol , 6*, 327-342.

Dunham, I., Shimizu, N., Roe, B. C., Hunt, A., Collins, J., Bruskiewich, R., et al. (1999). The DNA sequence of human chromosome 22. *Nature* , 402, 489-495.

Edman, P., & Begg, G. (1967). A protein sequenator. *Eur. J. Biochem , 1*, 80-91.

Edwards, N. (Performer). (2005). *Proteomics & Mass Spectrometry.* Maryland, College Park.

Eng, J. K., McCormack, A. L., & Yates, J. (1994). An approach to correlate tandem mass spectral data of peptides. *J. Am. Soc. Mass Spectrom , 5*, 976-989.

Fernández de Cossío, J., Gonzales, J., & Besada, V. (1995). A computer program to aid the sequencing of peptides in collision- activated decomposition experiments. *Comput. Appl. Biosci , 11*, 427–434.

Field, H., Fenyo, D., & Beavis, R. ,. (2002). A bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics , 2*, 36-47.

Filén, J. J. (2008). *Quantitative proteomics in the characterization of T helper lymphocyte differentiation.* Retrieved April 4, 2011, from http://www.doria.fi/bitstream/handle/10024/35952/D794.pdf?sequence=1

Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., et al. (2005). NovoHMM: A hidden markov model for *de novo* peptide sequencing. *Anal. Chem* , 77,7265-7273.

Flanagan, D. (1996). *Java in a nutshell: A desktop quick reference for java programmers* (Vol. 1). U.S.A.: O'Reilly & Associates Inc.

Forner, F., Foster, L. J., & Toppo, S. (2007). Mass spectrometry data analysis in the proteomics era. *Current Bioinformatics , 2*, 63-93.

Frank, A., & Pevzner, P. (2005). *De novo* peptide sequencing via probabilistic. *Anal. Chem* , 77, 964-973.

Frank, A., Savitski, M., Nielsen, M., Zubarev, R., & Pevzner, P. (2007). *De novo* peptide sequencing and identification with precision mass. *J. Proteome Res* , 6,114-123.

Godoy, L., Olsen, J. V., Souza, A. G., Li, G., Mortensen, P., & Mann, M. (2006). Status of complete proteome analysis by mass spectrometry:SILAC labeled yeast as a model system. *Genome Biology , 7* (6), R50.

Gosling, J., & Yellin, F. (1996.). *The Java Application Programming Interface* (Vol. 2). M.A.: Addison-Wesley.

Goto, M. A., & Schwabe, E. (2008). A Dynamic Programming Algorithm for *De Novo* Peptide Sequencing with Variable Scoring. *ISBRA , 4983*, 171-182.

Hamm, C., Wilson, W., & Harvan, D. (1986). Peptide sequencing program. Comput. Appl. *Biosci , 2*, 115–118.

Havilio, M. H., & Smilansky, Z. (2003). A intensity-based statistical scorer for tandem mass spectrometry. *Nal. Chem. , 75*, 435-444.

Henzel, W., Billeci, T., Stults, J., Wong, S., Grimley, C., & Watanabe, C. (1999). Identifying proteins from two-dimensional gels by molecular mass searching of peptide. *fragments in protein sequence databases , 90*, 5011-5015.

Hernandez, P., Müller, M., & Appel, R. D. (2006). Automated protein identification by tandem mass spectrometry: issues and strategies. *Mass Spectrometry Reviews , 25* (2), 235-254.

Hines, W., Falick, A., Burlingame, A., & Gibson, B. (1999). Pattern-based algorithm for peptide sequencing from tandem high-energy collision-induced dissociation mass-spectra. *J. Am. Soc. Mass Spectrom , 3*, 326–336.

Horn, D., Zubarev, R. A., & McLafferty, F. W. (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom , 11* (4), 320–332.

Huang, Y., Triscar, J., Tseng, G., Pasa-Tolic, L., Lipton, M., Smith, R., et al. (2005). Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal. Chem. , 77*, 5800-5813.

Ion source. (2005, June 27). Retrieved March 24, 2010, from http://ionsource.com/tutorial/isotopes/slide8.htm

Jakubowski. (2011, January 18). *The Central Dogma of Biology*. Retrieved April 10, 2011, from http://employees.csbsju.edu/hjakubowski/classes/ch331/dna/oldnacentdogma.html

Johnson, R. S., Martin, S. A., & Biemann, K. (1988). Collision-induced fragmentation of (M + H)+ ions of peptides. Side chain specific sequence ions. *International Journal of Mass Spectrometry and Ion Processes , 86*, 137-154.

Johnson, R., & Biemann, K. (1989). Computer-program (seqpep) to aid in the interpretation of high-energy collision tandem mass-spectra of peptides. *Biomed. Environ. Mass Spectrom , 18*, 945–957.

Keller, A., Purvine, S., Nesvizhskii, A., Stolyer, S., Goodlett, D., & Kolker, E. (2002). *Correct sequest search results using appended human peptide sequence database*. Retrieved April 17, 2011, from http://www.systemsbiology.org/extra/protein_mixture.html

Keller, A., Purvine, S., Nesvizhskii, A., Stolyer, S., Goodlett, D., & Kolker, E. (2002). Experimental protein mixture for validating tandem mass spectral analysis. *Omics , 6*, 207-212.

Kinter, M., & Sherman, E. N. (2000). *Protein sequencing and identification using tandem mass spectrometry.* New York: Wiley Interscience.

Krogh, A. (1998). *Guide to human genome computing* (2nd ed. b.). (M. J. Bishop, Dü.) San Diego.

Küng, S., Ongen, L., & Large, S. (2011, January 21). *TortoiseSVN.* Retrieved April 13, 2011, from TortoiseSVN web site: http://tortoisesvn.net/docs/release/TortoiseSVN_en/tsvn-intro-features.html

Large, S., Küng, S., & Onken, L. (2011, January 21). *TortoiseSVN.* Retrieved April 13, 2011, from TortoiseSVN web page: http://tortoisesvn.net/docs/release/TortoiseSVN_en/tsvn-introduction.html

Leipzig, J., Pevzner, P., & Hever, S. (2004). The alternative splicing gallery: Bridging the gap between genome and transcriptome. *Nucleic Acids Res.* , 32, 3977-3983.

Lemay, L., & Perkins, C. L. (1996). *Teach yourself Java in 21 days.* Sams Indianapolis, IN, USA: Sams.net.

Lu, B., & Chen, T. (2003). A suboptimal algorithm for *de novo* peptide sequencing via tandem mass spectrometry. *J Comput Biol , 10*, 1-12.

Lu, B., & Chen, T. (2004). Algorithms for *de novo* peptide sequencing via tandem mass spectrometry. *BIOSILICO , 2* (2), 85-90.

Ma, B., Doherty-Kirby, A., & Lajoie, G. (2003). PEAKS: Powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom , 17* (20), 2337-2342.

Mo, L., Dutta, D., Wan, Y., & Chen, T. (2007). MSNovo: A Dynamic Programming Algorithm for *de Novo* Peptide Sequencing via Tandem Mass Spectrometry. *Anal. Chem. , 79*, 4870-4878.

NetBeans IDE 6.5 Realise Information. (2008). Retrieved April 12, 2011, from NetBeans IDE 6.5 Realise Information website: http://netbeans.org/community/releases/65/index.html

NetBeans IDE 6.8 Release Information. (2009). Retrieved April 12, 2011, from NetBeans IDE 6.8 Release Information web site: http://netbeans.org/community/releases/68/

Netbeans IDE 6.9 Features. (2010). Retrieved April 12, 2011, from Netbeans 6.9 Feartures web site: http://netbeans.org/features/java/swing.html

NetBeans.org Community News. (2007, October 11). Retrieved April 12, 2011, from NetBeans.org Community News website: http://netbeans.org/community/news/show/1129.html

Niemeyer, P., & Peck, J. (1996). *Exploring Java* (Vol. 1). O'Reilly Media.

Origo. (2007). Retrieved April 13, 2011, from Origo Web site: http://origo.origo.ethz.ch/wiki/doc#What_is_Origo

Pandey, A., & Mann, M. (2000). Proteomics to study genes and genomes. *Nature , 405*, 837-846.

Papayannopoulos, I. (1995). The interpretation of collision-induced dissociation. *Mass Spectrometry Reviews , 14*, 49-73.

Peptide fragmentation. (2010). Retrieved March 18, 2011, from http://www.matrixscience.com/help/fragmentation_help.html

Perkins, D. N., Pappin, D., Creasy, D., & Cottrell, J. (1999). Probability-based protein identification by searching sequence. *Electrophoresis , 20* (18), 3551-3567.

Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., & Zhang, X. (2006). Performance evaluation of existing *de novo* sequencing. *J. Proteome Res. , 5*, 3018-3028.

Sakurai, T., Matsuo, T., Matsuda, H., & Katakuse, I. (1984). PAAS 3, a computer program to determine probable sequence of peptides from mass spectrometric data. *Biomed.Mass Spectrom , 11*, 396–399.

Santos-Rosa, H., Schneider, R., Bannister, A., Sherriff, J., Bernstein, B., Emre, N., et al. (2002). Active genes are tri-methylated at K4 of histone H3. *Nature* , 419, 407-411.

Scoble, H. A., Billet, J. E., & Biemann, K. (1987). A graphics display-oriented strategy for the. *amino-acid sequencing of peptides by tandem mass-spectrometry , 327*, 239–245.

Tanner, S., Shu, H., Frank, A., Mumby, M., Pevzner, P., & Bafna, V. (2005). Inspect:Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal. Chem* , 77, 4626-4639.

Taylor, J., & Johnson, R. (2001). Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal Chem , 73*, 2594-604.

Taylor, J., & Johnson, R. (1997). Sequence database searches via *de novo* peptide sequencing. *Rapid Commun Mass Spectrom , 11*, 1067-75.

Tulach, J. (2010). *Happy Birthday NetBeans -- interview with Jaroslav "Yarda" Tulach*. Retrieved April 12, 2011, from NetBeans web site: http://netbeans.org/community/articles/interviews/yarda-tulach.html

Varki, A., Cummings, R., Esko, J., Freeze, H., Hart, G., & Marth, J. (1999). *Essentials of glycobiology.* La Jolla, California: Cold Springs Harbor Laboratory Press.

Wan, Y., Yang, A., & Chen, T. (2006). PepHMM: A hidden markov model based scoring function. *Anal Chem , 78*, 432-437.

Washburn, M., Wolters, D., & Yates, J. (2001). Large scale analysis of the. *Nature Biotech- , 19*, 242-247.

Welcome to the NetBeans Community. (2008). Retrieved April 12, 2011, from netbeans web site: http://netbeans.org/about/index.html

Wells, J., & McLuckey, S. (2005). Collision-induced dissociation (CID) of peptides and proteins. *Meth. Enzymol. , 402*, 148-85.

Wilkins, M., Sanchez, J., Gooley, A., Appel, R., Humphery-Smith, I., Hochstrasser, D., et al. (1996). Progress with proteome projects:Why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. , 13*, 19-50.

Zaidi, K. (2006). *U.S.PHARMACOPEIA*. Retrieved March 12, 2011, from pharmacopeia.cn web site: http://www.pharmacopeia.cn/v29240/usp29nf24s0_c736.html

Zhang, N., Aebersold, R., & Schwikowski, B. (2002). ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics , 2* (10), 1406-1412.

Zhang, Z. (2004). *De novo* peptide sequencing based on a divide-and-conquer algorithm peptide tandem spectrum simulation. *Anal Chem , 76*, 6374-83.