RAY: A PROFILE-BASED APPROACH FOR HOMOLOGY MATCHING OF TANDEM-MS SPECTRA TO SEQUENCE DATABASES

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 Biotechnology

by
Şule YILMAZ

July 2012
İZMİR

We approve the thesis of **Şule YILMAZ**

Examining Committee Members:

Assoc. Prof. Dr. Jens ALLMER
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Assoc. Prof. Dr. Bilge KARAÇALI
Department of Electrical and Electronics Engineering, İzmir Institute of Technology

Assoc. Prof. Dr. Talat YALÇIN
Department of Chemistry, İzmir Institute of Technology

Assoc. Prof. Dr. Ahmet KOÇ
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Assist. Prof. Dr. Bünyamin AKGÜL
Department of Molecular Biology and Genetics, İzmir Institute of Technology

5 July 2012

Assoc. Prof. Dr. Jens ALLMER
Supervisor, Department of Molecular Biology and Gentics, İzmir Institute of Technology

Assoc. Prof. Dr. Bilge KARAÇALI
Co-Supervisor, Department of Electrical and Electronics Engineering, İzmir Institute of Technology

Assoc. Prof. Dr. Volga BULMUŞ
Head of the Department of Biotechnology and Bioengineering

Prof. Dr. R. Tuğrul SENGER
Dean of the Graduate School of Engineering and Sciences

ACKNOWLEDGMENTS

I would like to thank to my supervisor, Assoc. Prof. Dr. Jens ALLMER. He showed me how to be good in science, bioinformatics, academia and even Capoeira. He gave me a chance to work on computational mass spectrometry which I had no idea before and which eventually I really like it. His knowledge, wisdom, helps, comments and criticism are priceless on my works. I would like to express my deepest gratitude to him because he believed in me, motivated, supported and helped me to achieve my success for all the time. I am very grateful and I will always be. Working with him is one of my milestones in my career.

I would like to thank to Assoc. Prof. Dr. Bilge KARAÇALI for being my co-supervisor and also one of the committee members. I would like to express my thanks to Assoc. Prof. Dr. Talat YALÇIN, Assoc. Prof. Dr. Ahmet KOÇ and Assist. Prof. Dr. Bünyamin AKGÜL for being committee members.

I also would like to thank to my undergrad adviser from Bioengineering Department of Ege University, Prof. Dr. Erdal BEDİR for his supports every time.

I would like to thank to all my colleagues in İYTE – Bioinformatics. I would express my deepest thankfulness to İYTE – Mass Spectrometry Lab, especially to Assoc. Prof. Dr. Talat YALÇIN, Çağdaş TAŞOĞLU and Melike DİNÇ.

I would like to express my deepest thanks to my close friends, especially to Nihan AYDIN, Sibel DEMİROĞLU, Seçil SÖZER, Belgin AYTUN, Ceren AKKAN, Alperen GÜRGENÇ, Duygu SAÇAR and Hacer IŞIKTAŞ. It is great to know that they will be always with me in good times to celebrate with me and in bad times to support me.

I would like to express my sincere thanks to my family. Without them, none of these would have never ever happened and I will always appreciate them for their love, supports, and believes.

I would like to thank to TÜBİTAK for supporting this project and my study with scholarship.

ABSTRACT

RAY: A PROFILE-BASED APPROACH FOR HOMOLOGY MATCHING OF TANDEM-MS SPECTRA TO SEQUENCE DATABASES

Mass spectrometry is a tool that is commonly used in proteomics to identify and quantify proteins. Thousands of spectra can be obtained in just few hours. Computational methods enable the analysis of high-throughput studies. There are mainly two strategies: database search and *de novo* sequencing. Most of the researchers prefer database search as a first choice but any slight changes on protein can prevent identification. In such cases, *de novo* sequencing can be used. However, this approach highly depends on spectral quality and it is difficult to achieve predictions with full length sequence. Peptide sequence tags (PST) allows some flexibility on database searches. A PST is a short amino acid sequence with certain mass information but obtaining accurate PST is still arduous. In case a sequence is missing in database, homology searches can be useful. There are some homology search algorithms such as MS-BLAST, MS-Shotgun, FASTS. But, they are altered versions of existing algorithms, for example BLAST has been modified for mass spectrometric data and became MS-BLAST. Besides, they are usually coupled with *de novo* sequencing which still possess limitations. Therefore, there is a need for novel algorithms in order to increase the scope of homology searches. For this purpose, a novel approach that is based on sequence profiles has been implemented. A sequence profile is like a table that contains frequencies of all possible amino acids on a given MS/MS spectrum. Then, they are aligned to sequences in database. Profiles are more specific than PSTs and the requirement for precursor mass restrictions or enzyme information can be removed.

ÖZET

RAY: DİZİ VERİTABANLARINDA TANDEM MS SPEKTRALARIN HOMOLJİ EŞLEŞMESİNİ SAĞLAMAK AMACI İLE PROFİLE DAYALI YAKLAŞIM

Kütle spektrometresi, proteinleri tanımlamak ve miktarını belirmede Proteomiks’te sıkça kullanılan bir araçtır. Birkaç saat içinde, binlerce spektra elde etmek mümkündür. Bilişimsel yöntemler, böylesi yüksek verimliğe sahip çalışmalar hakkında bilgi alınmasını sağlar. Temel olarak iki strateji mevcuttur: veritabanı araması ve *de novo* sekanslama. Çoğu araştırmacının ilk tercihi veri tabanı aramasıdır; fakat proteindeki en ufak değişiklik bile tanımlanmayı engeller. Böylesi durumlarda, *de novo* sekanslama kullanılabilir. Fakat, bu yaklaşım spektral kaliteye oldukça bağlı olup, bütün dizi tahminlerini elde etmek oldukça zordur. Veritabanı aramalarına esneklik vermesi amacı ile, peptit dizi etiketleri (PDE) kullanılabilir. PDE belirli kütle bilgisini içeren kısa amino asit dizisidir, fakat hassas PDE’leri elde etmek hala güçtür. Bu nedenle, eğer veritabanında dizi bulunmuyorsa, homoloji aramaları yararlı olabilir. Günümüzde, bazı homoloji arama algoritmaları bulunmaktadır, MS-BLAST, MS-Shotgun, FASTS gibi. Fakat bunlar varolan algoritmaların değiştirilmiş halleridir, mesela BLAST kütle spektrometrik verisine göre modifiye edilmiş ve MS-BLAST adı verilmiştir. Ayrıca, bu algoritmalar, hala kısıtlamaları bulunan *de novo* sekanslama algoritmalarına dayalı çalışmaktadır. Homoloji aramalarının kapsamını genişletmek amacı ile yeni algoritmalara ihtiyaç duyulmaktadır. Bu amaçla, sekans profiline dayalı yeni bir yaklaşıma uygulanmıştır. Sekans profili, verilen MS/MS spektrumunda yer alan olası tüm amino asitlerin frekansını içeren bir tablodur. Elde edilen bu profiller sonra veritabanındaki dizilerle hizalanır. Profiller, PDE’lerden daha spesifiktirler ve öncül iyon kütlesine yada enzim bilgisine dair herhangi bir kısıtlama bulunmamaktadır.

To my father, Recep and my mother, Aysel

RAy

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

LIST OF ABBREVIATIONS

ACN	**:**Accession number
CID	**:** Collision-induced dissociation
ESI	**:** Electro spray ionization
FDR	**:** False discovery rate
HPLC	**:** High performance liquid chromatography
IT	**:** Ion trap
LC	**:** Liquid chromatography
MALDI	**:** Matrix assisted laser desorption ionization
MASSS	**:** Mass adjusted sequence similarity score
MS	**:** Mass spectrometry
MS/MS	**:** Tandem mass spectrometry
MS2	**:** Tandem mass spectrometry
MS-BLAST	**:** Mass spectrometry driven basic local alignment search tool
MS^n	**:** Multistage mass spectrometry
ORPR	**:** Overall removed peak ratio
PMF	**:** Peptide mass fingerprint
PST	**:** Peptide sequence tag
PTM	**:** Post translational modification
RDCR	**:** Removed doubly charged peak ratio
RPR	**:** Retained peak number ratio
TIC	**:**Total ion current

CHAPTER 1

INTRODUCTION

1. 1. Proteomics

Proteins are biochemical molecules that perform biological functions such as regulation (e.g. enzyme), immunological response (e.g. antibody), and transportation (e.g. hemoglobin) in any cells. They are composed of several polypeptides that consist of amino acids. The protein compositions vary in different cells, also at particular time and conditions. There is correlation between amino acid sequences and genomes and physiology which makes it possible to find out genomics information via amino acid sequences (Domon & Aebersold, 2006).

The proteome is the entire set of the proteins that are expressed by the genome at a particular time and under a certain condition. In addition, it may contain alternatively spliced or modified proteins (Forner, Foster, & Toppo, 2007; Allmer, 2011). Proteomics is the study of the proteome to identify proteins and also post translational modifications (Shadforth, Crowther, & Bessant, 2005), to map protein interactions (Cox & Mann, 2011), to characterize components of proteins and pathways in cells (Mann, Hendrickson, & Pandey, 2001) and to quantitate proteins (Aebersold & Mann, 2003).

Edman degradation (Edman, 1950) was used to sequence amino acids until 90's. But Edman-based analysis was usually slow due to sequencing of each peptide peak separately which prevents high-throughput proteomics studies (Gevaert & Vandekerckhove, 2000). In addition, this method usually failed to obtain long or accurate peptide sequences in case of any acetylations of amino terminus (Steen & Mann, 2004). Therefore, there was a need for new techniques in protein studies. Hence, the mass spectrometry technique took the place of Edman degradation due to sensitivity and speed features (Steen & Mann, 2004; Domon & Aebersold, 2006).

1. 2. Mass Spectrometry

Mass spectrometry is an analytical technique that measures mass-to-charge (m/z) ratios of analytes in a gas phase. A mass spectrometer consists of three main components which are an ion source, an analyzer and a detector, respectively (Figure 1.1). An analyte, which is shown in Figure 1.1 as input – a sample mixture, is entered into an ion source where it becomes ionized depending on different ionization techniques. Then, the ionized sample passes through an analyzer which separates the ions in a sample according to their m/z ratios by electrical field under vacuum. Finally, the ions hit a detector which records number of ions at each m/z value and then a mass spectrum is generated. In addition, mass spectrometric analysis for protein are usually done with the positive mode on where the ions are protonated instead of the negative mode which is deduction of protons from peptide (Mann, Hendrickson, & Pandey, 2001).

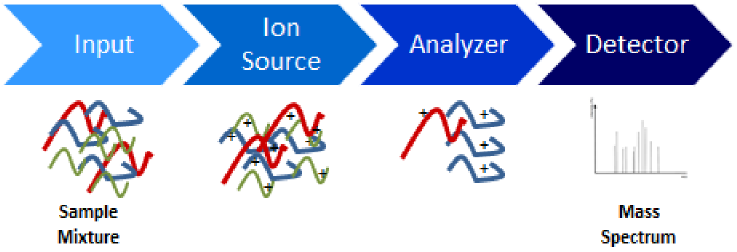


Figure 1.1. Main mass spectrometry components: *ion source, analyzer, detector*. A sample mixture is ionized in ion source then ions are separated through an analyzer according to their m/z ratios. Number of ions is recorded via a detector and finally a mass spectrum is generated.

Mass spectrometry has been studied over decades but in the late 80's two ionization techniques which are electro-spray ionization (ESI) and matrix assisted laser desorption ionization (MALDI) allow ionizing of thermally instable proteins (Domon & Aebersold, 2006). Since then, mass spectrometry has become essential in proteomics studies (Aebersold & Mann, 2003). The studies of John Bennett Fenn and Koichi Tanaka who developed these ionization techniques for biological compounds were awarded the Nobel Prize in Chemistry, 2002. They are soft-ionization techniques that allow the formation of ions from entire protein into a gas phase with little fragmentations (Kiner & Sherman, 2000; Tuli & Ressom, 2009)

Mass spectrometry can be used with several stages to get information about proteins under study. A single-stage mass spectrometry (MS), which is shown in Figure 1.1, enables measuring peptide fragments in a protein sample and it forms a “peptide mass fingerprint (PMF)”. A PMF makes it possible to assign proteins with 2 peptides and also speeds up database searches (Henzel, Watanabe, & Stults, 2003). On the other hand, any changes on mass such as post-translational modifications or single nucleotide polymorphisms halt predictions and secondly database size affects protein predictions (McHugh & Arthur, 2008). Thirdly, PMF performs well for a single protein, but not protein mixtures (Olsen & Mann, 2004). In addition to MS, tandem mass spectrometry can be used to determine the primary structure, attachment sites or post translational modifications. In this approach, particular ions from a single stage are selected and then fragmented by collisions with an inert gas in a collision cell (“fragmentation unit” in Figure 1.2.) between two analyzers inside of a mass spectrometer machine. Since two analyzers are used, the technique is called “tandem mass spectrometry (MS/MS or MS2)” (Steen & Mann, 2004; Domon & Aebersold, 2006) (Figure 1.2). A fragmented peptide is called a “precursor ion” or a “parent ion” and the resultant ions measured in the sequential analyzers are named as “product ions” or “daughter ions” (Standing, 2003). If a fragment is without any charges, it is referred to as “neutral loss” and cannot be detected directly. So, to improve identification especially for complex sample mixtures, further fragmentation sections are performed in additional stages and this method is termed “multistage mass spectrometry (MS^n)” (Olsen & Mann, 2004; Steen & Mann, 2004; Bandeira, Olsen, Mann, & Pevzner, 2008).

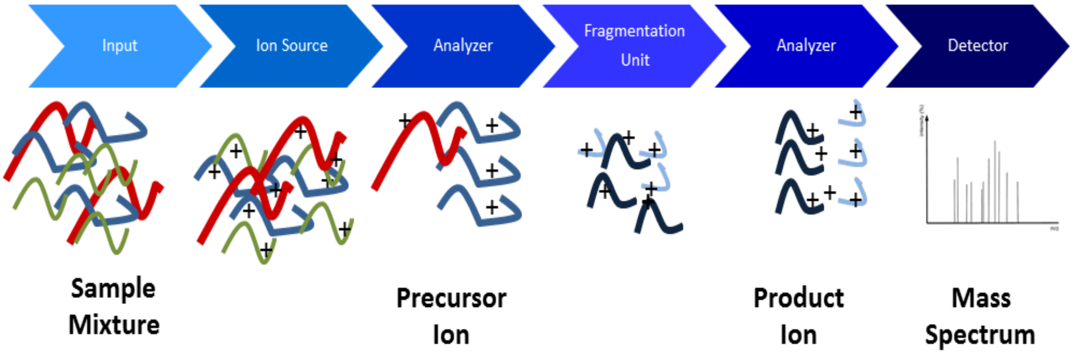


Figure 1.2. Tandem mass spectrometry (MS/MS). A sample mixture is ionized at an ion source and then ions are separated through an analyzer according to their m/z values. After the first analyzer, some ions are selected (“precursor ion”) and they are fragmented in a fragmentation unit so as to obtain segment peptides (“product ion”). Then they are separated throughout the second analyzer. At the end of the second analyzer, number of ions hit to a detector is recorded and a MS/MS spectrum is generated.

In MS-based proteomics, MALDI and ESI are commonly used as ion sources. In MALDI, peptides are charged singly predominantly whereas in ESI there are mainly multiply charged ions (Aebersold & Goodlett, 2001). Besides, quadrupole (Q), time-of-flight (TOF) and ion traps (especially in MS^3 studies) are commonly used analyzers in this field (Steen & Mann, 2004).

1. 3. Bottom-up and Top-down Proteomics

Proteomics strategies with mass spectrometry can be divided into two main groups: *bottom-up* and *top-down* (Figure 1.3). In bottom-up proteomics, proteins are digested with proteases such as trypsin prior to MS analysis. Then, peptide fragments are ionized and transformed into a gas phase without any fragmentations. This technique is appropriate for peptide identification, but not ideal for any modifications (Chait, 2006).

In top-down proteomics, a sample is directly introduced to mass spectrometry. Then, fragmentation occurs inside of a mass spectrometer. This approach enables studying modifications in addition to peptide identification (Chait, 2006). Nevertheless, there are limitation for protein size (<50 kDa) and dynamic range (Kelleher, 2004).

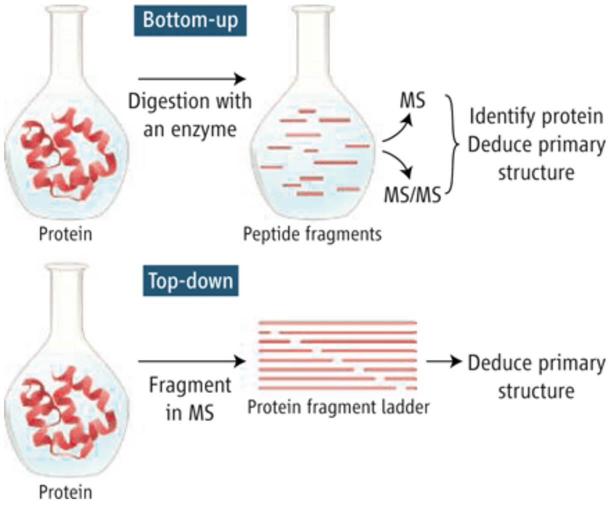


Figure 1.3. *Bottom-up* and *top-down* proteomics. In bottom-up strategy, a sample is digested with enzyme before introducing to a mass spectrometer. In top-down approach, an intact protein is ionized and converted into gas phase then fragmentation occurs inside of a mass spectrometer (Source: Chait, 2006).

1. 4. Peptide Fragmentation

Peptide fragmentation enables getting the primary structure information of peptide under study. It is possible to ionize entire peptides due to soft-ionization feature of MS and then they can be selected as precursor ions which are fragmented into product ions in a fragmentation unit (Figure 1.2) by exposing low energy (<100eV) on the molecule. There are several activation techniques for fragmentation (Eidhammer, Flikka, Martens, & Mikalsen, 2008) which are in-source decay (ISD), post-source decay (PSD), laser-induced dissociation (LID), collision-induced dissociation (CID, or also referred to as collisionally-activated dissociated, CAD), electron transfer dissociation (ETD), and electron capture dissociation (ECD).

This phenomenon occurs in a predictable way and the accepted naming is the Roepstorff-Fohlmann-Biemann nomenclature which was explained according to fragmentation status with CID (Forner, Foster, & Toppo, 2007). In theory, fragmentation can occur at any bonds on peptide (Allmer, 2011). In regard to the “mobile proton model”, protons move along peptide backbone and while increasing dissociation energy protonation weakens amide bond which leads fragmentation at this bond (Eidhammer, Flikka, Martens, & Mikalsen, 2008; Forner, Foster & Troppo, 2007; Seidler, Zinn, Boehm, & Lehmann, 2010). As a result of this fragmentation, two different prevailing ion types are obtained: *b* and *y* ion respect to charge status whether in amino-terminus and carboxyl-terminus respectively. These two ions are complementary to each other. If fragment ions are in amino-terminus, they are labeled *a*, *b* and *c* and the complementing fragment ions in carboxyl-terminus are named *x*, *y* and *z* (Figure 1.4.). *b* ions usually have satellite ions which are lower 28 u due to neutral loss of carbon monoxide and actually *a* ions (Aebersold & Goodlett, 2001). Depending on MS analyzer, prevailing ion types vary such as for quadrupole instrument *y* ions are predominant , but for ion traps *b* and *y* ions are (Forner, Foster, & Toppo, 2007; Steen & Mann, 2004) Some undesired fragmentations can happen such as an internal fragmentation (Cottrell, 2011). Moreover, side chain loss (scl) can also be observed, especially with high collision energy (>500 eV) (Johnson, Davis, Taylor, & Patterson, 2005). If side chain loss fragmentation takes place in *a*-, *y*- and *z*-ions, scI ions are labeled *d*-, *v*- and *w*-ions (Eidhammer, Flikka, Martens, & Mikalsen, 2008). In addition, subscripts in fragment ions show number of side chains groups (R) in that region. For example in Figure 1.4, a_2 ion contains 2 side chain groups (R_1 and R_2) and a_2 is complementary to x_{n-1} which holds actually n-1 side chain groups.

A fragmentation ladder is a group of peptide fragments. They share the same terminus information, but are different from each other due to length and each element is apart from each other by one amino acid (Aebersold & Goodlett, 2001; Chait, 2006; Forner, Foster & Troppo, 2007). An example of a consecutive fragment ladder would be b_3, b_4, b_5 and b_6. They are the same ion types that have terminus in common which is the amino-terminus in this case.

Peptide fragmentation success depends on the abundance the of precursor ion, fragmentation amount and fragmentation energy (Allmer, 2011). Hence, fragmentation may not be always successful.

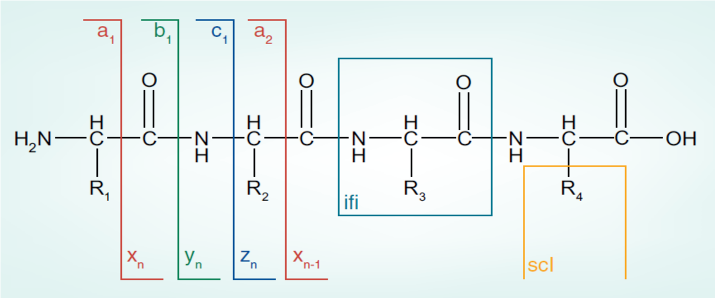


Figure 1.4. Peptide fragmentation. Proteins are usually fragmented on peptide bonds and *a, b* and *c* ions are fragment in amino terminus and *x, y* and *z* ions are complementary ions in carboxyl terminus part. Subscripts show location of ions. ifi: internal fragment ions and scI: side chain lose (Source: Allmer, 2011).

1. 5. Deisotoping and Charge State Deconvolution

Preprocessing MS/MS spectra allows more accurate and specific peptide identification. For this purpose, some preprocesses such as spectrum denoising, precursor ion charge state recognition, calibration and centroiding are implemented (Gentzel, Köcher, Ponnusamy, & Wilm, 2003; Forner, Foster & Troppo, 2007; Eidhammer, Flikka, Martens, & Mikalsen, 2008).

An isotope is one of several versions of one chemical element which has the same atomic number (number of protons), but differs in mass number (number of neutrons and protons). There are several isotopes of atoms such as carbon (C), hydrogen (H), oxygen (O), nitrogen (N) and also sulfur (S) which occur in the composition of biological molecules. They have different natural abundances. For instance, the carbon atom has 16 isotopes including Carbon-12, Carbon-13 and Carbon-14. Due to different natural abundances of the isotopes for different atoms, isotopic envelopes are observed in a MS/MS spectrum. They are like a cluster of peaks that are separated by small m/z differences due to mass differences of the isotopes. Peptide charges enables determination of the distance between isotopic peaks in an envelope, for instance the distances equals to 1 Da and 0.5 Da for the singly and doubly charged peaks respectively (Matthiesen, 2006). Monoisotopic mass is the mass value of the most abundant isotopic ions per elements whereas the average mass is the weighted sum of all isotopes for an element based on their natural abundance (Eidhammer, Flikka, Martens, & Mikalsen, 2008). Isotopomer are isotopic isomers which composed of the same elements, but different locations. Isotopic envelopes include monoisotopic peaks and isotopomers and therefore, they are also called an “isotopomeric envelope” (Hellerstein & Neese, 1999) (Figure 1.5).

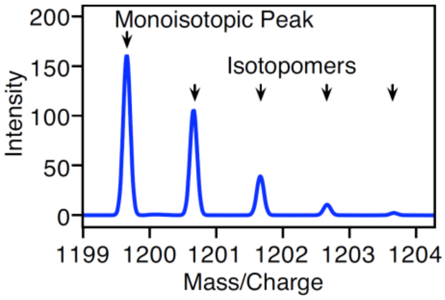


Figure 1.5. An isotopic envelope which is composed of several isotopomers
(Source: Sykes & Williamson, 2008).

Deisotoping is the process that aims to reduce isotopic envelopes into a single peak. This process is generally followed by another process. Herein ions with various charges (up to three or four, but not higher than charge of precursor mass) are converted to singly charged status and intensities are summed up if the translated ion already exists on a mass spectrum. This step is termed “charge state deconvolution”. Deisotoping and charge state deconvolution are the processes that take place concurrently. As a result of these two steps, the complexity level of mass spectra is reduced since there are fewer fragments that are represented by more than one datum.

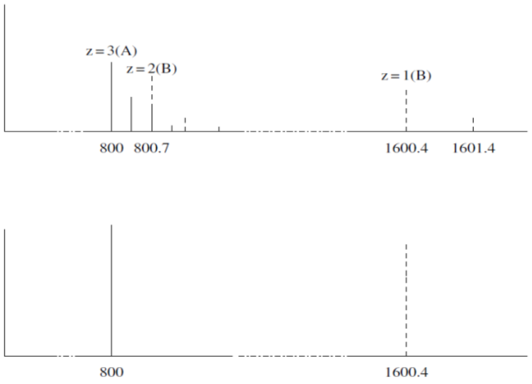


Figure 1.6. Deisotoping and charge state deconvolution. One fragment ion (A) with triply charged is shown as full line and it has four isotopic peaks. Another fragment ion (B) is indicated as dashed line and doubly and singly charged B ions have three and two isotopic peaks respectively. There are also overlapping isotopic envelopes for both ions. After deisotoping and deconvolution, intensities are summed up (Source: Eidhammer, Flikka, Martens, & Mikalsen, 2008).

1. 6. Computational Methods

Tandem mass spectrometry (MS/MS) is one of the tools in protein identification. Correctly identified proteins are crucial in proteomics experiments. To identify peptides, there are basically two methods: database search and *de novo* sequencing. Database search algorithms depend on sequences in the databases whereas *de novo* sequencing can find a peptide without any aids from a database, directly from a tandem MS spectrum under study (Allmer, 2011).

1.6.1. Database Search

Database search algorithms are the first choice of many researchers in proteomics (Xu & Ma, 2006). The algorithms differ from each other in terms of their algorithmic aspects and scoring functions. Sequest (Eng, Mccormack, & Yates, 1994) and Mascot (Perkins, Pappin, Creasy, & Cottrell, 1999) are the commercial software. OMSSA (Geer et al., 2004), X!Tandem (Craig & Beavis, 2004) and MyriMatch (Tabb, Fernando, & Chambers, 2007) are some of the freely available tools.

Protein sequences in databases are digested by a user-defined enzyme, and then theoretical spectra for the fragmented peptides which have similar masses to precursor mass of the experimental MS/MS spectrum under study are created. Likelihoods are examined by scoring functions that are specific to algorithms (Figure 1.7). There are probability and non-probability based scoring functions (Forner, Foster & Troppo, 2007).Hits with high scores may not be the correct sequence. Therefore, statistical analysis such as e-value or p-value must be carried out (Eidhammer, Flikka, Martens, & Mikalsen, 2008).

Database search algorithms depend on correctly annotated sequences in the database. It fails in case of unexpected post-translational modifications, novel proteins, lack of genomic data, incorrect predictions and alternative splicing (Xu & Ma, 2006; Allmer, 2011).

Mass tolerance and enzyme information are some of the search parameters that influence the results dramatically. Mass tolerance shows the correlation between an experimental and a theoretical spectrum. It is related to the mass accuracy of the instrument (Forner , Foster & Troppo, 2007) and mass accuracy varies with different mass analyzers.

Besides, sensitivity is a limiting factor in predictions that discriminates the correct results (true positive) from incorrect results (false positive). For this purpose, false discovery rates (FDR) are calculated. This statistical term is used to explain a proportion of incorrect predictions which have lower scores than thresholds within search parameters (Elias & Gygi, 2010). There are two approaches to evaluate FDR (Elias & Gygi, 2007): *target-decoy* and *target|decoy*. Targets are the sequences found in the organism under the study whereas decoys are shuffled version of targets to indicate incorrect hits. While creation of decoy sequences, some features such as similar length, similar amino acid compositions and uncommon peptides between target and decoy sequences can be considered. In *target-decoy* approach, two sequences are merged into one database. In *target|decoy* strategy, two separate databases are used.

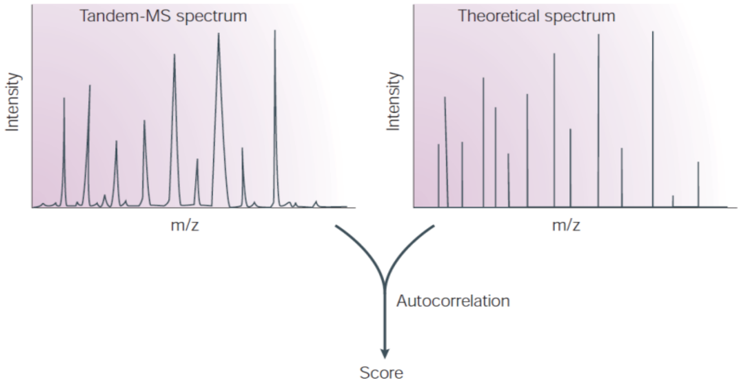


Figure 1.7. Database search. Sequences in a database are virtually digested if precursor mass of MS/MS spectrum matches to mass of those sequences. That is followed by generation of theoretical spectra. Then an experimental MS/MS spectrum and a theoretical MS/MS spectrum are compared by scoring function (Source: Steen & Mann, 2004).

1.6.1.1. Peptide Sequence Tag

A peptide sequence tag (PST) is a short amino acid sequence with certain mass information that can be found in an experimental MS/MS spectrum (Mann & Wilm, 1994). They are like a signature of a peptide and usually of a 3-amino acid length. They are divided into three regions, m_1 (from 0 to the start m/z of the first amino acid in the sequence), the partial sequence, and m_3 (from the last m/z of the last amino acid in the sequence to precursor mass). The mass information of the partial sequence is known as well. For even noisy and incomplete fragmentations in some spectra, there are some parts that can be identified partially instead of a full construction. This was the starting point of the PST study of Mann *et al.* (Mann & Wilm, 1994). This approach gives flexibility to database searches especially for modified proteins.

There are several PST algorithms, for example DirecTag (Tabb, Ma, Martin, Ham, & Chambers, 2008), GutenTag (Tabb, Saraf, & Yates, 2003), InsPect (Tanner, Shu, Frank, Wang, Zandi, Mumby, Pevzner & Bafna, 2005) and MultiTag (Sunyaev, Liska, Golod, Shevchenko, & Shevchenko, 2003). The algorithms usually work based on singly charged fragment ions but they should consider that some ions especially in higher charged spectra are multiply charged (Sun, Zhang, & Liu, 2011). Besides, they have different aspect in algorithms. For instance, GutenTag considers all ions in a PST are *y* ions and use *b* ions are used to confirm *y* ions (Tabb, Saraf, & Yates, 2003).

They can be used to speed up database search (Frank, Tanner, Bafna, & Pevzner, 2005), to identify peptides and also PTMs (Cao & Nesvizhskii, 2008) and to assess spectral quality (Ham, Aerni, Cheek, Whitwell, Caprioli, Tabb & Ma, 2011).

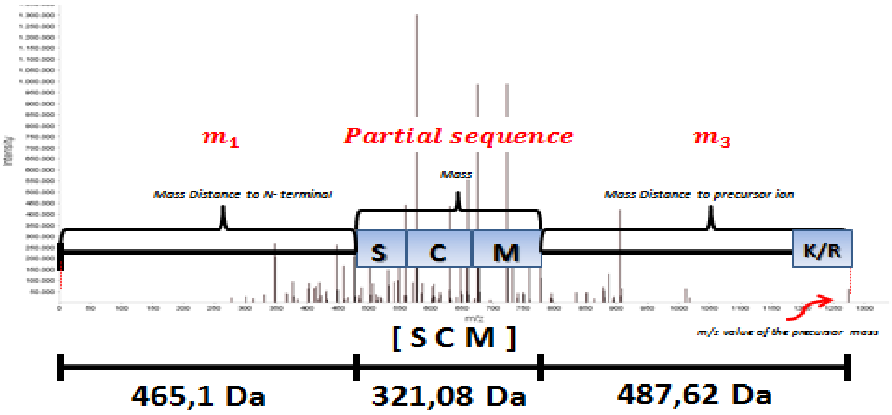

m_1
Partial sequence
m_3
Mass Distance to N-terminal
Mass
Mass Distance to precursor ion
S
C
M
K/R
Intensity
m/z
m/z value of the precursor mass
[S C M]
465,1 Da
321,08 Da
487,62 Da

Figure 1.8. A peptide sequence tag (PST). A PST is a short amino acid sequence with unique mass information. They are specific to an experimental MS/MS spectrum and used in database searches.

1.6.2. *De Novo* Sequencing

De novo sequencing algorithms work on directly experimental MS/MS spectra without any aids from databases. Database searches fail if a protein in study is novel or a modified version of a known protein (Xu & Ma, 2006), alternatively spliced (Allmer, 2011), or genome of the organism has not been sequenced yet (Shadforth, Crowther, Bessant, 2005). In such cases, de *novo* sequencing can be useful to identify peptides. The strategy is based on the fact that proteins are fragmented in a predictable manner (See Section 1.4. Peptide Fragmentation) (Forner, Foster & Troppo, 2007). In addition to identification, they can be used to validate database results and homology-based searches (Xu & Ma, 2006). However, *de novo* sequencing predictions depend on spectral quality and precision of mass spectrometer (McHugh & Arthur, 2008).

There are several *de novo* sequencing algorithms. Some of them are PepNovo (Frank & Pevzner, 2005), Lutefisk (Taylor & Johnson, 1997), SHERENGA (Dancík, Addona, Clauser, Vath, & Pevzner, 1999), PEAKS (Ma, Zhang, Hendrie, Liang, Li, Doherty-Kirby & Lajoie, 2003).

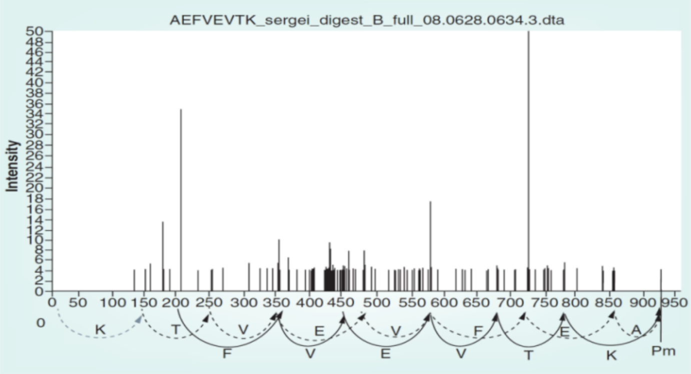


Figure 1.9. *De novo* sequencing. The algorithms work on directly spectrum itself without any aids from databases to sequence a peptide (Source: Allmer, 2011).

1.6.3. Homology Search

Homology search can be useful to identify peptides if the sequence is missing in database (Shevchenko, Sunyaev, Loboda, Bork, Ens, & Standing, 2001) since homologous proteins have some identical peptides in common. BLAST, FASTA and Shotgun programs were modified to work on mass spectrometric data and became MS-BLAST (Shevchenko, Sunyaev, Loboda, Bork, Ens, & Standing, 2001), FASTS (Mackey, Haystead, & Pearson, 2001) and MS-Shotgun (Huang, Jacob, Pegg, Baldwin, Wang, Burlingame & Babbitt, 2001) in homology based studies. A general procedure in these homology-based tools starts with getting *de novo* sequencing results and then carrying out similarity searches to get hits from a database. The problem with these tools is underestimating the *de novo* sequencing error (Xu & Ma, 2006). To overcome this problem, other homology-based programs were introduced, such as OpenSea (Searle, Dasari, Turner, Reddy, Choi, Wilmarth Mccormack, David, Nagalla, 2004) and Spider (Han, Ma, Zhang, & Na, 2004). OpenSea algorithm is based on mass information, not amino acid codes. The algorithm works on tag creation and then breadth-first search. OpenSea considers *de novo* sequencing errors. Nevertheless, it does not allow *de novo* sequencing errors and homolog mutations at the same position. This problem is overcome by another homology-based search tool named Spider.

1.6.4. Spectral Profile

Currently, there are problems in existing protein identification tools. For example, even though extensive studies have been carried out in *de novo* sequencing, it is still difficult to achieve full length of peptide sequences and also percentage of fully reconstructed peptide sequences is less than 50% (Kim, Bandeira, & Pevzner, 2009). Besides, local qualities in MS/MS spectra prevent identifications. Herein, local quality means that some part of an experimental MS/MS spectrum is lack of consecutive ions in fragmentation ladder to construct a full peptide sequence. Therefore, there must be some novel methods in protein identification.

Spectral profiles (Kim, Bandeira, & Pevzner, 2009) are used to represent MS/MS spectra with probability of all possible amino acids (range is [0-1]). They are similar to “motif profile” in bioinformatics, but here the information is not known for sure. In addition, they may look like “scored spectra”, nevertheless they work globally instead of a local satellite peaks.

Kim *et al.* used spectral profiles to create gapped peptide sequences. Within a gapped peptide sequence, some ambiguous regions caused by local quality are shown as between brackets with mass values. This approach enables short and/or long gaps on a spectrum. They can be considered as a niche between full-length *de novo* sequences and PSTs (Kim et al., 2009).

Gapped peptide sequences are used in database searches to find homologs. In addition to homology searches, they also speed up database searches. They provide information even poor quality spectra as well. Gapped peptides give more accurate results compared to PSTs.

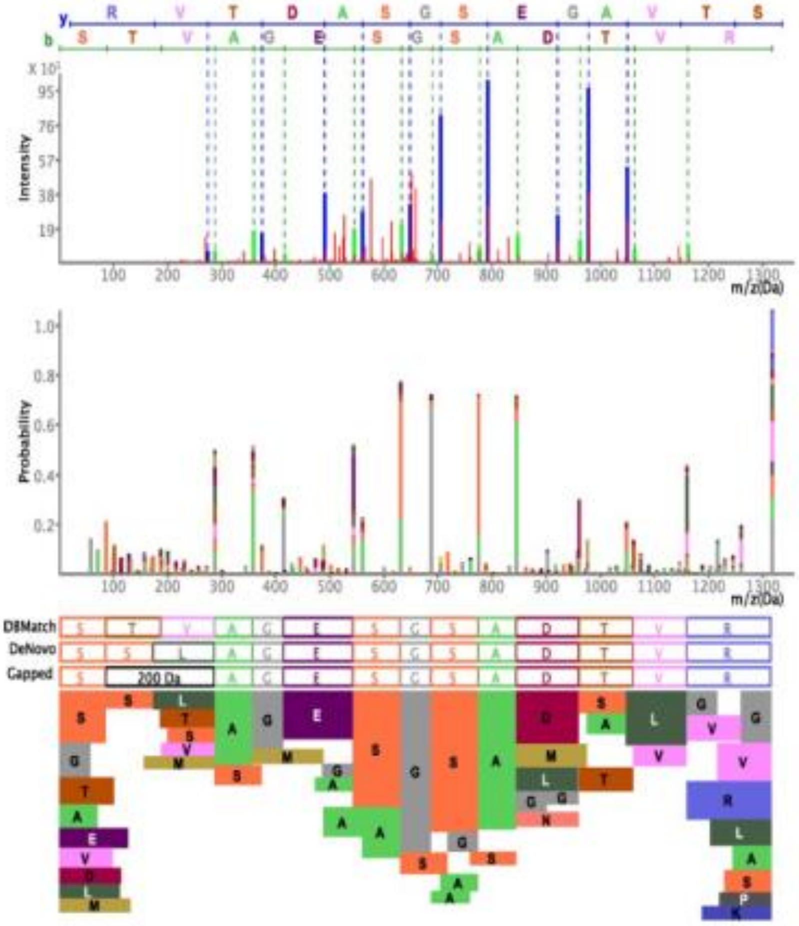


Figure 1.10. A spectral profile. Top: An experimental spectrum with *b*-ions (green) and *y*-ions (blue). Middle: A spectral profile shows probability of all possible amino acids in an experimental spectrum. Bottom: database match (DBMatch), *de novo* prediction (DeNovo) and gapped peptide sequences (Gapped) respectively. A gapped peptide is created by a spectral profile and ambiguous regions are shown between brackets (Source: Kim, Bandeira, & Pevzner, 2009).

1.7. The Aim of the Study

Computational methods for mass spectrometry-based proteomics enables to identify proteins under study. Database search is the most commonly used approach by researches, but slight changes on protein sequences prevent identification. *De novo* sequencing is another approach which works without any aids from a database, but it is still difficult to achieve full length protein prediction in spite of extensive studies in the recent years.

Homology searches can be useful to get information about the protein in study when the sequence is missing in database. In such cases, getting information about those homologous proteins can be useful due to homologous proteins have some peptides in common. Existing homology search tools are generally coupled with *de novo* sequencing algorithms. The results from those homology searches depend on the success of *de novo* algorithms which owns some handicaps and does not give prediction successfully. Also, they are generally modified versions of another bioinformatics tool, for example MS-BLAST is modified version of BLAST so as to work with mass spectrometric data.

Therefore, new strategies are necessary in order to increase the scope of homology searches for comprehensive sequence databases. For this purpose, we are presenting a new algorithm that is based on a novel approach. Herein, sequence profiles are constructed for a given MS/MS spectrum and then they are aligned to sequences in a database via the Smith-Waterman algorithm. A profile is specific to a MS/MS spectrum and contains frequencies of all possible amino acids that can be found on a given MS/MS spectrum. They are like a table where each row shows amino acids whereas each column indicates frequency at a sequential position.

CHAPTER 2

MATERIAL AND METHODS

2.1. Tandem Mass Spectra Data Sets

2.1.1. Keller *et al.* dataset

The published dataset (Keller *et al.*, 2002) is composed of two protein mixtures (A and B) which are cleaved by trypsin. In mixtures there are 18 proteins (bovine beta-casein, bovine carbonic anhydrase, bovine cytochrome c, bovine beta-lactoglobulin, bovine alpha-lactalbumin, bovine serum albumin, chick ovalbumin, bovine transferrin, rabbit gapdh, rabbit phosphorylase b, *E.coli* beta-galactosidase, bovine gamma-actin, bovine catalase, rabbit myosin, *E.coli* alkaline phosphatase, horse myoglobin, *B.lichenformis* alpha-amylase, *S.cerevisiae* phosphomannose isomerase). MS analysis is done with LCQ tandem mass spectrometer which liquid chromatography (LC) is coupled MS/MS with ESI-ITMS (TheromoFinnigan, San Jose, CA). Some spectra are selected and in the selected dataset there are 109 spectra with singly charged precursor ions, 629 spectra with doubly charged precursor ions and 18 spectra with triply charged precursor ion.

2.1.2. Synthetic Peptide Dataset

There are two synthetic peptide datasets. The first synthetic peptide dataset is composed of 4 different peptide sequences (ASCMGLY, AVFDRKSDAK, CLGGLLTMV and GLCTLVAML). It contains 20 MS/MS spectra of the synthetic peptides with singly charged precursor mass in this dataset. The second one is composed of 45 peptides that are derived from cytochrome c (ACN P00004), bovine serum albumin (ACN P02769), ovalalbumin (ACN P01012), myoglobin (ACN P68082) and lysozyme C (ACN P61626).

The first and the second datasets are measured with AB Applied Biosystems MDS SCIEX 4000 ESI Q-Trap and Thermo Scientific LTQ XL Linear Ion Trap ESI mass spectrometers with CID fragmentation respectively. Prepared peptide mixture is introduced to a mass spectrometer directly, without any prior chromatography steps. Herein, parameters including filling time, activation time, collision energy (CID), TIC and cycle numbers are changed to obtain spectra with different spectral quality regarding to existence of different ion types.

2.1.3. Synthetic Spectra Dataset

The dataset contains 32 synthetic spectra which are different regarding to *b*-ions, *y*-ions and *a*-ions. The ions for the expected sequences are calculated via tool named MS-Product (UCSF, by Burlingame). In addition, there are 110 spectra with *a*-, *b*- and *y*- ions.

2.2. Databases

Three databases are used to evaluate performance of this algorithm.

- Synthetic spectra databases: Entries which are all expected and modified (added/deleted/mutated/shuffled) sequences.
- Keller *et al.* – control mixture database: 103 entries which all protein sequences in the dataset and also human keratin sequences
- Synthetic peptide databases: Uniprot databases for chicken (9031), bos taurus, human (9606) and horse (9796) which include expected and modified sequences.
- Non-redundant database

2.3. RAy Algorithm

The algorithm has two steps: *profile construction* and *profile alignment* (Figure 2.1). In the first step, profiles are constructed per each MS/MS spectrum (See Section

2.3.1. Profile Construction). Firstly, a MS/MS spectrum is preprocessed and any possible amino acids (fit amino acids) are found out which is followed by score maximization. Then, frequencies of all possible amino acids are calculated per location. In the second step, profiles are aligned against to sequences in database by using Smith-Waterman algorithm (See Section 2.3.2. Profile Alignment).

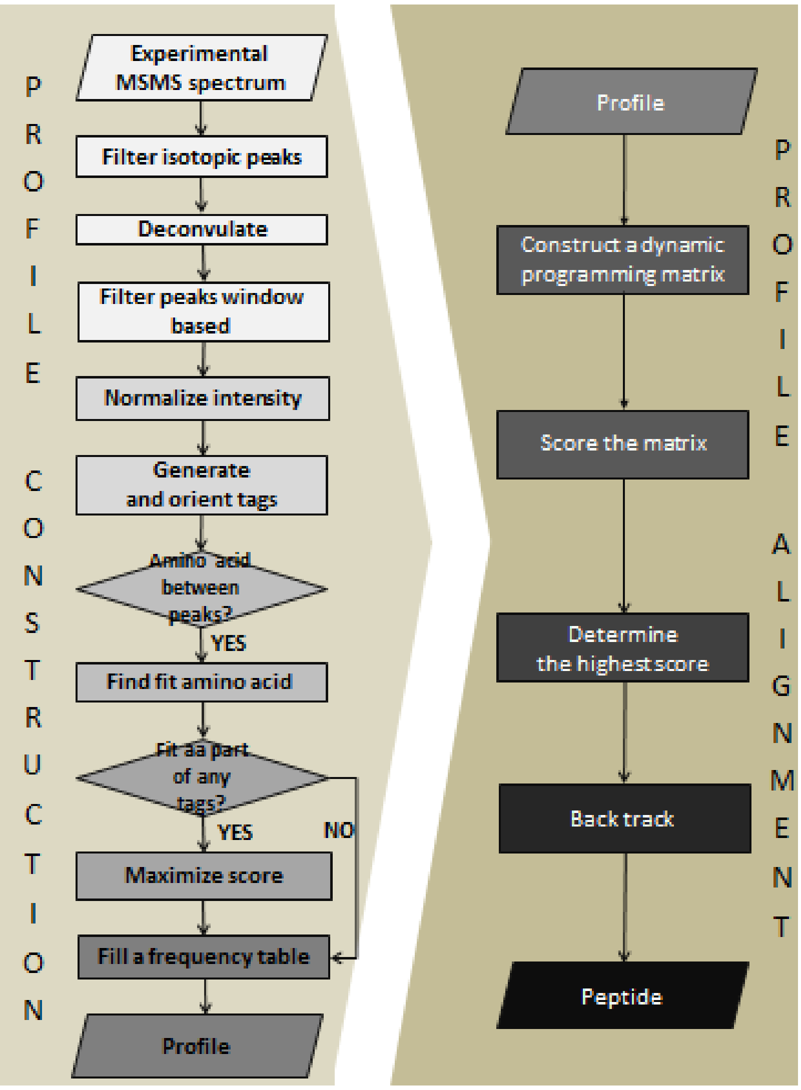


Figure 2.1. The flowchart of RAy. In the first step of RAy, a MS/MS spectrum specific profile is constructed and then aligned against to sequences in a database.

The algorithm is implemented in JAVA. There are some libraries used during the implementation which are Massspeclib, Seqlib and Helpers. Massspeclib and Seqlib libraries are created by Jens Allmer and developing by Allmer’s group in İYTE. Massspeclib library provides MassSpectrum, MSTypeFactory, Filtering, Tag classes that enable working on mass spectrometric data. Seqlib library provides sequence related classes such as Amino Acid, Amino Acids for protein sequences and make it possible to read any database files such as FASTA. Helpers class facilitates doing statistical and also edit distance score calculations. In addition, Junit4 and JUnit3 testing are run.

2.3.1. Profile Construction

2.3.1.1. Preprocessing Experimental Spectra

2.3.1.1.1. Isotopic Peaks Filtration & Deconvolution

The first step in preprocessing is collapse isotopic peaks (See Section 1.5.Deisotoping and Charge State Deconvolution). Herein, the isotopic envelopes are collapsed into one peak.

The second step in preprocessing is deconvolution (Figure 2.2). First of all $m/_z$ value of doubly charged precursor ion (PM/2) is determined. Then, existence of any doubly charged ions from zero to $m/_{z_{PM^{2+}}}$ values are checked in case that any corresponding singly charged ions are found. Herein, intensity of singly charged peak must be higher than assumed doubly charged peaks. Within the doubly charged ions, standard deviation (*std)* and average (*avg)* values of intensities are calculated. This step enables determination of the intensity limits for deconvolution. After that, the ions which have intensities are within the limits are assumed as doubly charged and then the charge state is converted into one according to Equation 2.1. If there are no singly charged ions for the converted ions prior to this step, the doubly charged peak is directly translated into charge state one (Light gray in Figure 2.2). In case that there is already singly charged ion in that location, the intensity of converted ion is added to the current singly charged ion (Dark gray in Figure 2.2).

$$({m/_z}^{2+} * 2) - 1 = {m/_z}^{+} \tag{2.1}$$

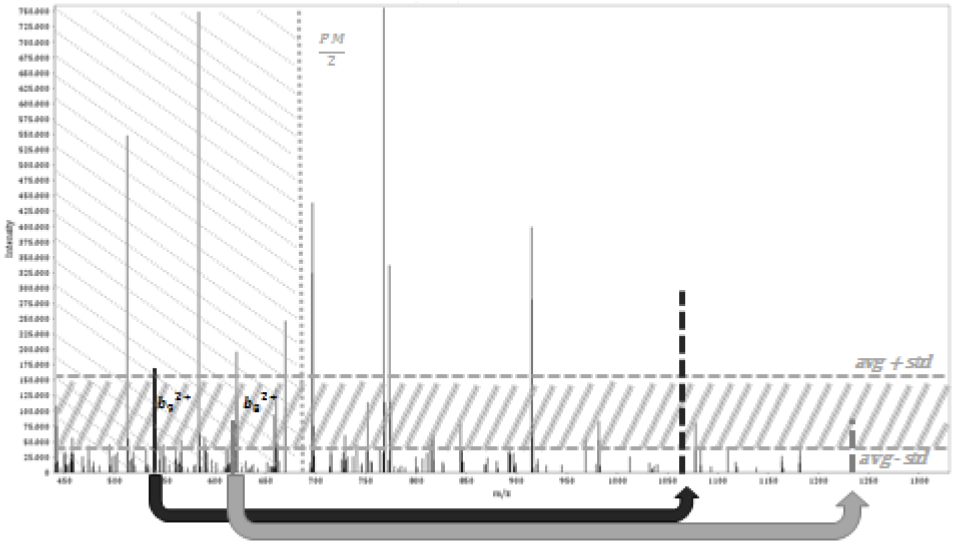


Figure 2.2. Deconvolution. Any peaks that have lower $m/_z$ values than doubly charged precursor ions can be doubly charged. Limits are determined for charge state deconvolution by checking existence of any doubly charged ions and then their average (*avg)* and standard deviation (*std)* of intensities in that location are calculated. Peaks that have intensity values between (*avg-std)* and *(avg+std)* are assumed charge status with two and then they are deconvulated into one. If the singly charged peak already exists, intensity is increased. Otherwise, a translated peak with charge state one is introduce to a MS/MS spectrum.

2.3.1.1.2. Window Based Peak Filtration

Then window based filtering is carried out. For this purpose, $(Precursor\ mass/50)$ windows are created for an experimental MS/MS spectrum and top 7 peaks that are sorted by intensity values are selected for each window (Figure 2.3).

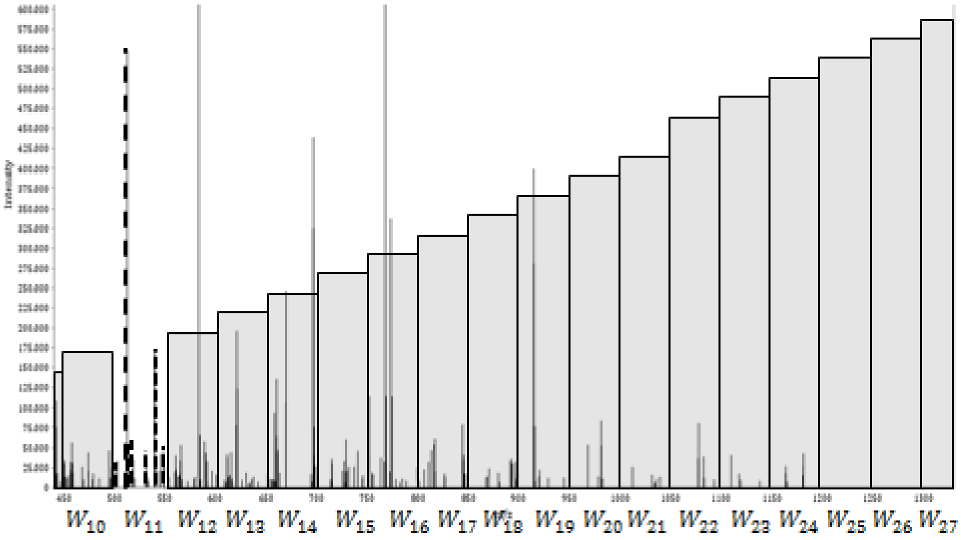


Figure 2.3. Window based filtering. An experimental MS/MS spectrum is divided into windows and then, the top 5 peaks with the highest intensities are selected. For instance, on the figure the dashed lines are the peaks with the highest values in the given window (W_{11}) therefore, they are selected.

2.3.1.1.3. Intensity Normalization

The last step in preprocessing is intensity normalization. All intensities are sorted and the top 5% of and the lowest 5% of intensities are discarded temporarily for statistical evaluation. If discarded numbers in the top are less than 3, the top 3 intensities and the bottom 3 intensities are removed. Then average (*avg)* and standard deviation (*std)* values are calculated from the rest. According to average and standard deviation limits, intensities for all peaks are classified into 6 different groups and their intensity values are updated (Table 2.1).

Table 2.1. Classification limits for intensity normalization

Maximum (Exclusive)	**Minimum (Inclusive)**	**Value**
+∞	*avg* + (2* *std*)	6
avg + (2 * *std*)	*avg* + *std*	5
avg + std	*Avg*	4
Avg	*avg – std*	3
avg – (*std*)	*avg* – (2**std*)	2
avg – (2**std*)	- ∞	1

2.3.1.2. Finding Fit Amino Acids

It is likely to find any amino acids between peaks on a MS/MS spectrum. They may be *b*-, *y*- or any other ions types (See Section 1.4. Peptide Fragmentation). RAy finds all possible amino acids referred as "*fit amino acids*". For example, there is a fit amino acid between two peaks (which are 342.10 and 489.20) that equals to mass of glutamic acid on the spectrum in Keller *et al*. dataset named GPFPII_sergei _digest_A_full _05.1518.1520.1.dta (Figure 2.4).

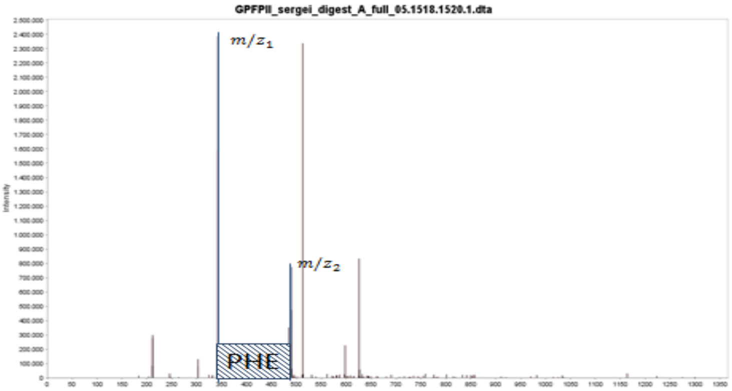


Figure 2.4. Finding fit amino acids. It is possible to find any amino acids between peaks in a tandem MS spectrum. Herein, differences between the peaks which have the m/z values are 342.10 and 489.20 respectively gives glutamic acid (Phe).

There is a parameter which enables finding out fit amino acids within certain mass differences. It is called as “mass tolerance”. Depending on the mass tolerance, mass accuracy that affects certainty of fit amino acids can be adjusted. In single step MS/MS analysis, leucine (L) and isoleucine (I) cannot be differentiated since their masses are the same. Therefore, they are shown as J. In addition if MS analysis is not accurate enough, glutamine (Q) and lysine (K) cannot be differentiated as well. In such cases, these fit amino acids are shown as B. Mass values of all amino acids and their one letter codes are listed in Table 2.2 (See Section 1.5.Deisotoping and Charge State Deconvolution for monoisotopic and average mass).

Table 2.2. Amino acid information

Amino acid name	**3 letter code**	**1 letter code**	**Precision adjusted**	**Monoisotopic mass**	**Avereage Mass**
Glycine	Gly	G	G	57.021417	57.052
Alanine	Ala	A	A	71.03712	71.079
Serine	Ser	S	S	87.03203	87.078
Proline	Pro	P	P	97.05277	97.117
Valine	Val	V	V	99.06842	99.133
Threonine	Thr	T	T	101.04768	101.105
Cysteine	Cys	C	C	103.00919	103.144
Isoleucine	Ile	I	**J**	113.08407	113.160
Leucine	Leu	L	**J**	113.08407	113.160

(**cont. on the next page)**

Table 2.2. (cont.)

Asparagine	Asn	N	N	114.04293	114.104
Aspartic acid	Asp	D	D	115.02695	115.089
Glutamine	Gln	Q	**B**	128.05858	128.131
Lysine	Lys	K	**B**	128.09497	128.174
Glutamic acid	Glu	E	E	129.04260	129.116
Methionine	Met	M	M	131.04049	131.198
Histidine	His	H	H	137.05891	137.142
Phenylalanine	Phe	F	F	147.06842	174.177
Arginine	Arg	R	R	156.10112	156.188
Tyrosine	Tyr	Y	Y	163.06333	163.170
Tryptophan	Try	W	W	186.07932	186.213

2.3.1.2.1. Score Maximization

A tag is a short sequence with unique mass information (See Section 1.6.1.2.Peptide Sequence Tag). Numerous tags can be generated on a one MS/MS spectrum. In order to obtain more an informative profile, tags are benefited. Because, any fit amino acids which take place on any tags are more likely to be part of the full length peptide sequence comparing to any individual fit amino acids.

In order to fill a frequency table, scores for each fit amino acid are calculated. To calculate a score for any fit amino acids, two sub-scores are considered which are related to intensity and tag. Firstly, the intensities of start and end m/z values are averaged. Then, in case that a fit amino acid is part of any tags, its score is increasing. Here, a tag score is calculated by multiplication of tag length with 2. At the end, these two values are summed up (Equation 2.2 and 2.3.).

$$Score_{fittingAA} = Avg_{intensity} + Score_{tag} \qquad (2.2)$$

$$Score_{fittingAA} = \frac{start_{\ intensity} + end_{intensity}}{2} + (len(tag) * 2) \qquad (2.3)$$

2.3.1.3. Orientation Elimination

It is not certain which part of an experimental MS/MS spectrum is a carboxyl or an amine terminus of peptide in the study. There are two approaches to eliminate orientation in this algorithm: *fit amino acid* and *tag* based.

First of all, any tags are generated with a mass tolerance related parameter named “tag precision”. If the precision is bigger than 0.5, it is set to 0.1 Da.

Current tag-based approaches do not overcome the orientation problem. For instance, GutenTag algorithm assumes that all ion series in a tag is y-ions (See Section 1.4.Peptide Fragmentation and 1.6.1.2.Peptide Sequence Tag). Herein, an algorithm specific approach is implemented in order to overcome that problem (Figure 2.5). Each tag is put into rows and the direction of the first tag.is assumed the correct direction. For instance, the tags with sequences of EAK, AVTH and RA are put into the first, second and third rows respectively. For each box in one row is different from each other with 300 Da. Windows are created based on mirror symmetry matched with column size of 2 (W_{a1} is matched to W_{b1} in Figure 2.5).Then, orientation check process starts from overlaps in the middle windows. In order to correct orientation, $Score_{orientation}$ is calculated (Equation 2.4). Here the most frequent amino acid is counted and the value is divided by sum of all frequencies in one column. If the row is reversed and $Score_{orientation}$ is increased as a result, direction is changed.

$$Score_{orientation} = \frac{Count_{the\ most\ frequent_{column}}}{Count\ _{all\ frequencies_{column}}} \tag{2.4}$$

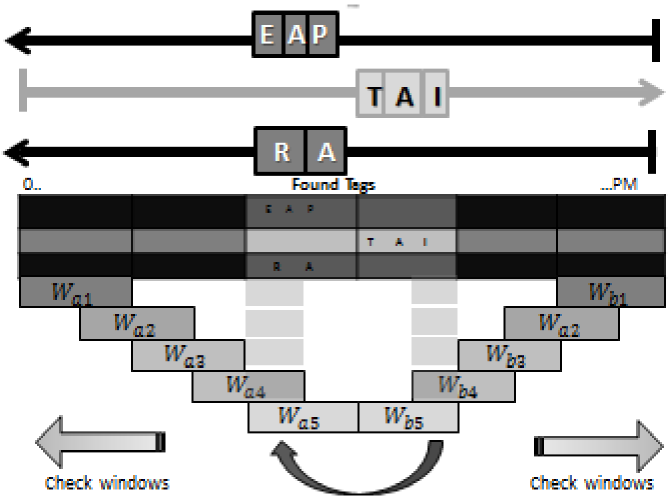
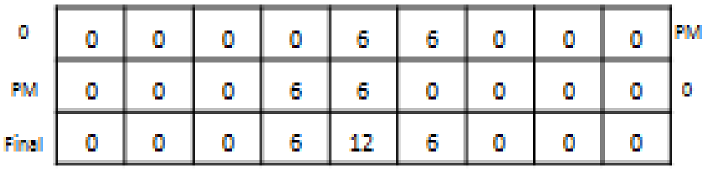


Figure 2.5. Tag orientation elimination.

After directionality fix in tags, score per each fit amino acid is calculated (Equation2.2) and a pre-frequency table is filled by these scores. In the beginning, each fit amino acid is converted to individual rows and each row has cells that contain mass information. Mass of each cell equals to glycine mass (Equation 2.5). Then, according to start and end $^{m}/_{z}$ values, a score is filled per locations. For instance, one glycine is a fit amino acid with the start and the end $^{m}/_{z}$ values are 238.08 and 285.1 respectively. Besides, $score_{fittingAA}$ equals to 6. From zero to precursor mass, the determined locations in respect to $^{m}/_{z}$ values are filled with 6. Then in order to eliminate orientation, filling starts from precursor mass to zero. If the score from this direction increases the existing score, the scores are summed up to obtain "final" row (Figure 2.6).

$$Mass\ Range_{profile} = \left(\frac{Monoisotopic\ Mass_{Precursor\ ion}}{Monoisotopic\ Mass_{Glycine}}\right) \qquad (2.5)$$



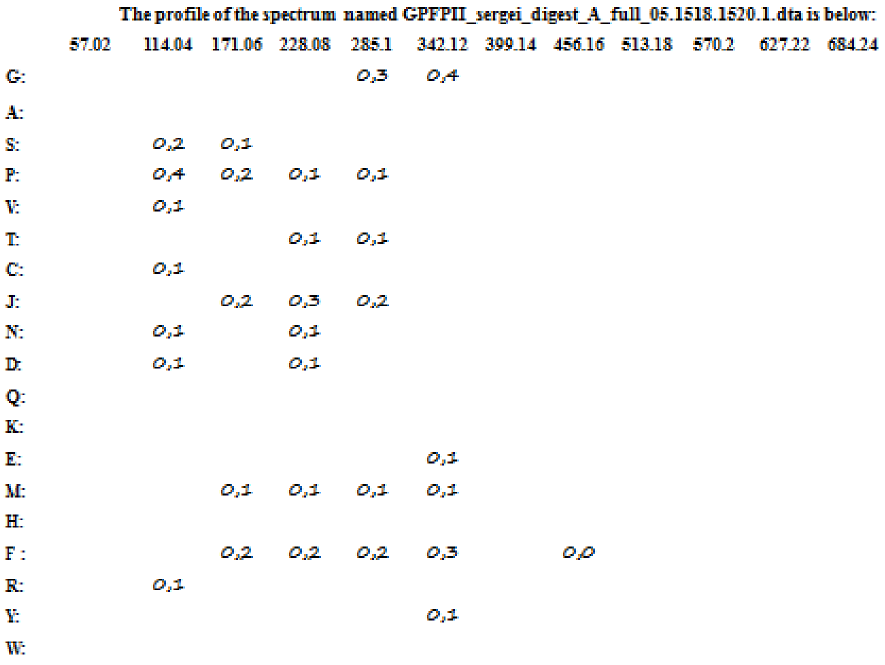
0	0	0	0	0	6	6	0	0	0	PM
PM	0	0	0	6	6	0	0	0	0	0
Final	0	0	0	6	12	6	0	0	0	

Figure 2.6. Filling principle. On the figure, a fit amino acid is glycine with the start and end $^{m}/_{z}$ values are 238.08 and 285.1 respectively. From zero to precursor mass, it is filling from the 5th cell till the 7th cell. Then from precursor mass to zero, scoring is at 4th and 5th cell. Since the score in 4th cell increases the existing score, the final score is summed of these two values (Note: indexing starts from zero instead of one).

2.3.1.4. Filling a Frequency Table and Obtaining a Profile

The last step is obtaining a profile from a frequency table. On a pre-frequency table, each row indicates a fit amino acid and each column shows locations without directionality. In order to get a sequence profile, the scores on the same column for each individual amino acid are summed up and then divided by overall scores on that column. As a result, frequency per each amino acid which varies from 0.0 to 1.0 is calculated.

A sequence profile is a table that each row represents certain amino acids and each column indicates positions without directionality information. Each cell contains frequencies. The Figure 2.7 is as an example for a profile. It is one of the spectra in Keller *et al.* dataset with singly charged precursor ion and the expected sequence is GPFPII.



```
              The profile of the spectrum  named GPFPII_sergei_digest_A_full_05.1518.1520.1.dta is below:
      57.02    114.04   171.06   228.08   285.1    342.12   399.14   456.16   513.18   570.2    627.22   684.24
G:                                        0,3      0,4
A:
S:             0,2      0,1
P:             0,4      0,2      0,1      0,1
V:             0,1
T:                               0,1      0,1
C:             0,1
J:                      0,2      0,3      0,2
N:             0,1               0,1
D:             0,1               0,1
Q:
K:
E:                                                 0,1
M:                      0,1      0,1      0,1      0,1
H:
F :                     0,2      0,2      0,2      0,3               0,0
R:             0,1
Y:                                                 0,1
W:
```

Figure 2.7. A sequence profile. This is one of the spectra from Keller *et al.* dataset. A sequence profile contains frequencies per each amino acid in columns. Each row indicates amino acid and each column shows location without directionality.

Parameters for profile constructions are summarized in Table 2.3.

Table 2.3. Parameters for the profile construction

Name	**Default**	**Information**
Mass tolerance	0.5 Da	To find out fit amino acids.
Tag precision	0.1 Da	To find out amino acids for a tag construction. Masses of the consecutive amino acids in a tag must be less than tag tolerance.
Times of standard deviation	1	To set limits (*avg- (times*stdv))* for deconvolution.
Window Num	PM/50	To set the window number while window based filtration.
Window Size	7	To set the peak number in a window during window based filtration.
Isotopic Peak Filtering	On	In order to process filtration based on isotopic envelopes collapsing.
Window Based Filtering	On	In order to process filtration based on selection of the top peaks regarding to intensity values for window based.

2.3.2. Profile Alignment

Profiles are aligned against sequences in a database by using Smith-Waterman algorithm which is one of dynamic programming algorithms (Smith & Waterman, 1981). This algorithm guarantees the optimal results depending on scoring systems. Generally speaking, a scoring matrix is created and then results are obtained via back track.

2.3.2.1. Construction of a Dynamic Programming Matrix

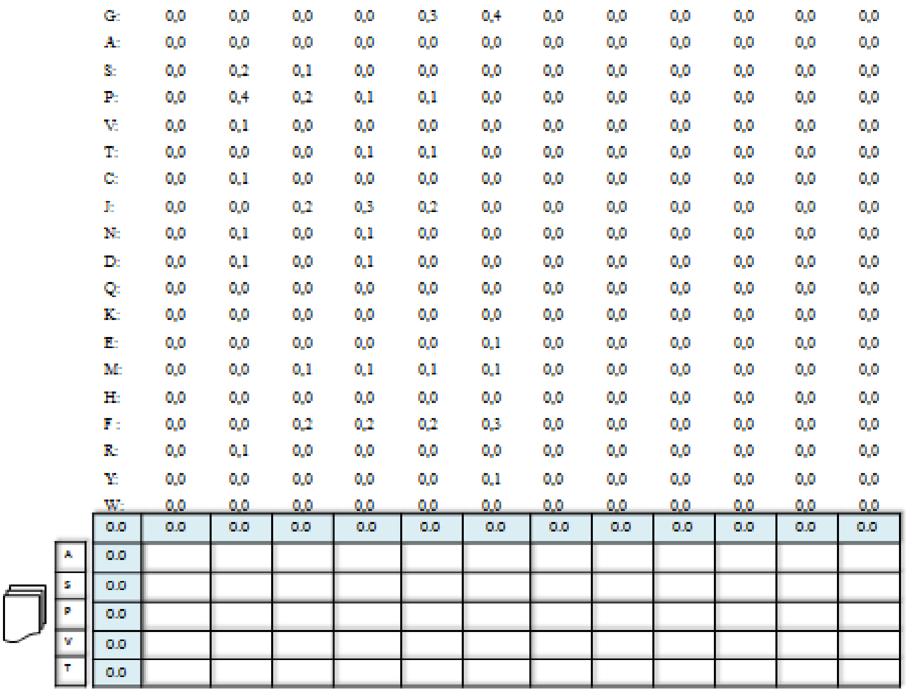
A dynamic programming matrix (M) is a table which contains scores between a sequence and a profile. It has one additional row and column filled with zero. For example, if a profile has 12 columns and a length of a sequence in database is 5, the matrix will be constructed with 12+1 columns and 5+1 rows (Figure 2.8).

$i \in columns\ of\ a\ profile$

$j \in characters\ in\ a\ sequence$

$$M[i][0] = 0.0$$

$$M[0][j] = 0.0$$



G:	0,0	0,0	0,0	0,0	0,3	0,4	0,0	0,0	0,0	0,0	0,0	0,0
A:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
S:	0,0	0,2	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
P:	0,0	0,4	0,2	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
V:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
T:	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
C:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
I:	0,0	0,0	0,2	0,3	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0
N:	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
D:	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Q:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
K:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
E:	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
M:	0,0	0,0	0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0
H:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
F :	0,0	0,0	0,2	0,2	0,2	0,3	0,0	0,0	0,0	0,0	0,0	0,0
R:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Y:	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
W:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0												
S	0.0												
P	0.0												
V	0.0												
T	0.0												

Figure 2.8. Construction of a dynamic programming matrix. It has one additional row and column filled up with zeros. Row number equals to (length of a sequence+1) and column number equals to (columns of a profile+1).

2.3.2.2. Scoring a Dynamic Programming Matrix

Scoring of the dynamic matrix is based on the rule (Equation 2.6). There are four possibilities for the score in a cell of a matrix. The highest value between these four options is selected for a value of a cell (Figure 2.9).

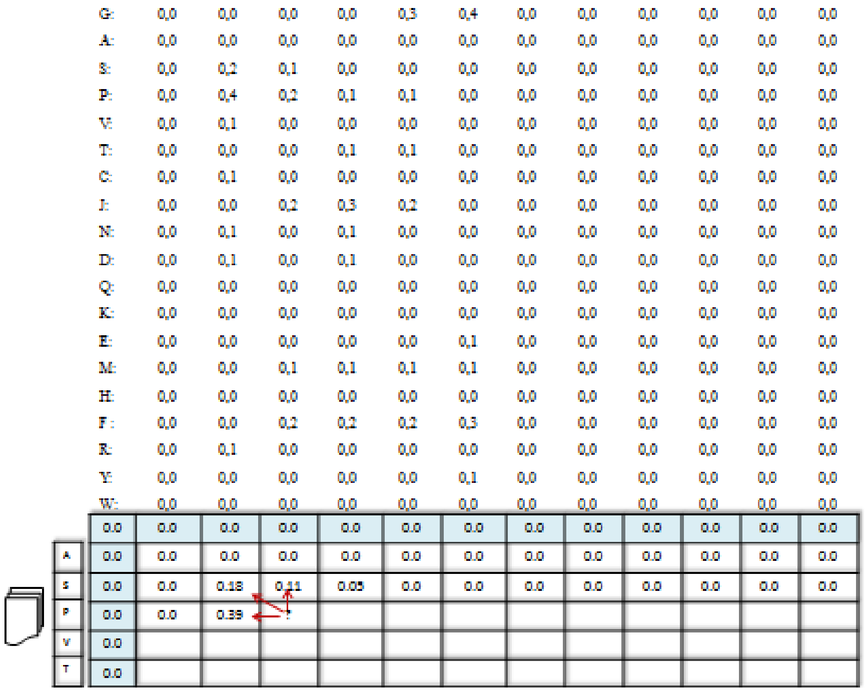
1. *Diagonal Score:* The sum of two scores which are a score from the cell at the upper left ($M[i-1][j-1]$) and a similarity score shown as $s(i,j)$. Two options are available to get a substitution score; using a simple substitution matrix and any known substitution matrices such as BLOSUM, PAM. For a simple substitution matrix, in case that an amino acid matches to itself, the score equals to a match score, otherwise a mismatch score. The two scores can be set by a user. The similarity scores are calculated by multiplication of a substitution score and a frequency of each amino acid on a profile. The substitution score is calculated per each amino acid *(a)* in a profile against one amino acid (b) of a sequence. Then, the highest similarity score is taken and that the information of amino acid in profile is selected (Equation 2.7).
2. *Left Score*: The score from the left of the current cell at the matrix ($M[i][j-1]$) is deducted by the gap score (g).

3. *Up Score:* The score from the up of the current cell ($M[i-1][j]$) is subtracted by the gap score (g).

4. *Zero.* This option prevents any negative values on a scoring matrix.

$$M[i][j] = \begin{cases} M[i-1][j-1] + s(i,j) \\ M[i][j-1] + g \\ M[i-1][j] + g \\ 0.0 \end{cases} \tag{2.6}$$

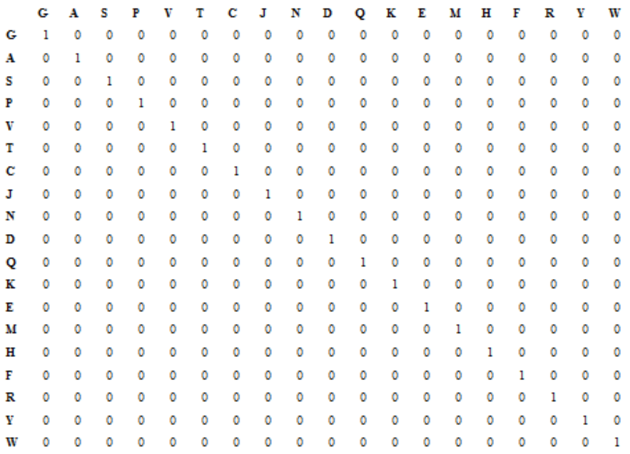
$$s(i,j) = frequency_a . substitution(a,b) \tag{2.7}$$



G:	0,0	0,0	0,0	0,0	0,3	0,4	0,0	0,0	0,0	0,0	0,0	0,0
A:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
S:	0,0	0,2	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
P:	0,0	0,4	0,2	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
V:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
T:	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
C:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
I:	0,0	0,0	0,2	0,3	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0
N:	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
D:	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Q:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
K:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
E:	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
M:	0,0	0,0	0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0
H:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
F :	0,0	0,0	0,2	0,2	0,2	0,3	0,0	0,0	0,0	0,0	0,0	0,0
R:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Y:	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
W:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
S	0.0	0.0	0.18	0.11	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	0.0	0.0	0.39	?									
V	0.0												
T	0.0												

Figure 2.9. Scoring a dynamic programming matrix. There are four possibilities to score any cells. These possibilities are indicated by arrows. If the score comes from upper left, similarity score is added to it. If the score is from left and up, the gap score is introduced to them. The fourth option, which is zero, prevents any negative values in the matrix.

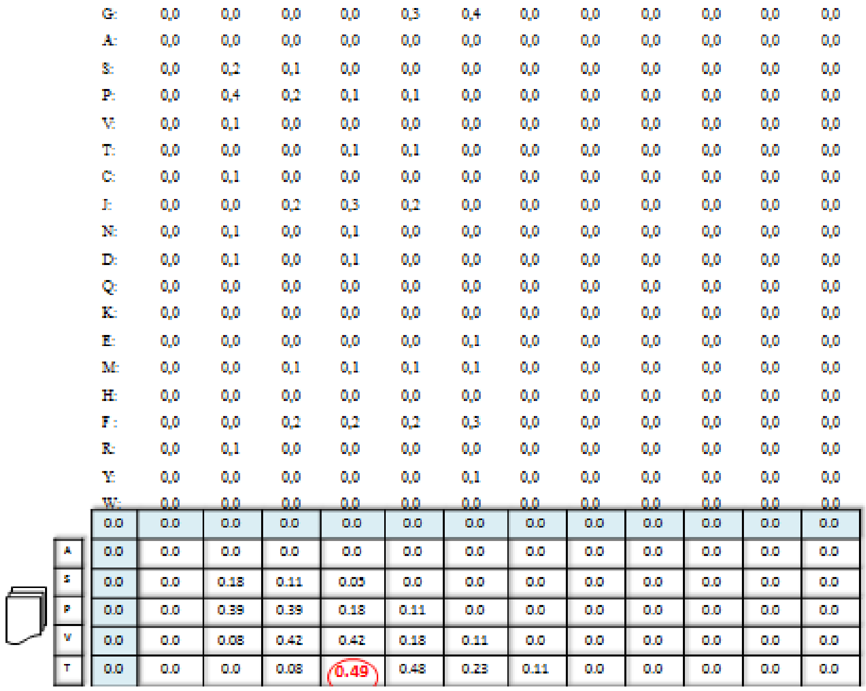


	G	A	S	P	V	T	C	J	N	D	Q	K	E	M	H	F	R	Y	W
G	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 2.10. A simple substitution matrix. If an amino acid matches itself, it gives 1 (match score); otherwise the substitution score is 0 (mismatch score). Moreover, known substitution matrices can be used (e.g. BLOSUM or PAM). These values can be set by a user.

2.3.2.3. Determination of the Highest Score

The next step after filling out the scoring matrix is determination of the highest score. It may be anywhere in the matrix. For instance, in the example of Figure 2.11, the highest score is 0.49 (which is circled).



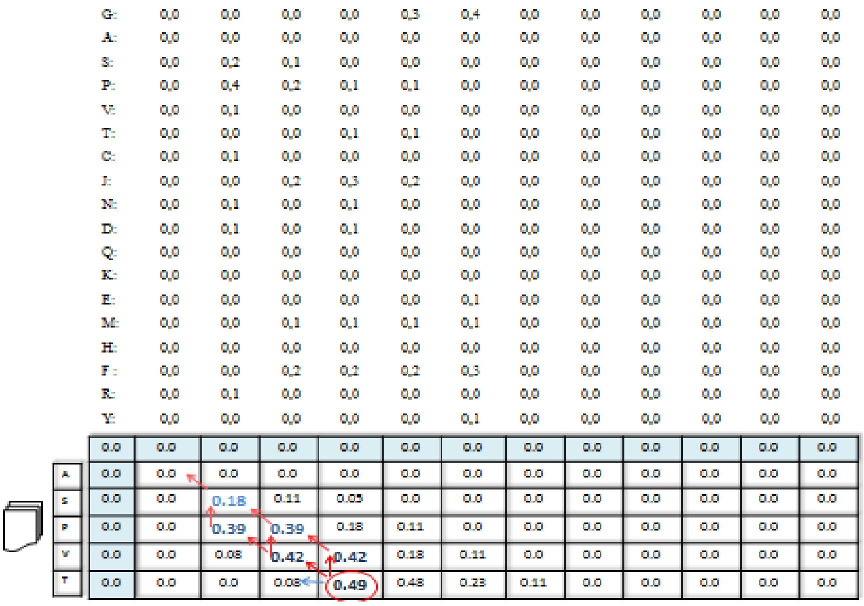
G:	0,0	0,0	0,0	0,0	0,3	0,4	0,0	0,0	0,0	0,0	0,0	0,0
A:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
S:	0,0	0,2	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
P:	0,0	0,4	0,2	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
V:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
T:	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
C:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
J:	0,0	0,0	0,2	0,3	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0
N:	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
D:	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Q:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
K:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
E:	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
M:	0,0	0,0	0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0
H:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
F :	0,0	0,0	0,2	0,2	0,2	0,3	0,0	0,0	0,0	0,0	0,0	0,0
R:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Y:	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
W:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
S	0.0	0.0	0.18	0.11	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	0.0	0.0	0.39	0.39	0.18	0.11	0.0	0.0	0.0	0.0	0.0	0.0	0.0
V	0.0	0.0	0.08	0.42	0.42	0.18	0.11	0.0	0.0	0.0	0.0	0.0	0.0
T	0.0	0.0	0.0	0.08	0.49	0.48	0.23	0.11	0.0	0.0	0.0	0.0	0.0

Figure 2.11. Determination of the highest score. The value can be anywhere in the matrix. For example here the highest score is 0.49 which is located in $M[3][5]$. (Note: indexing starts from zero, not one).

2.3.2.4. Back Track

Back track starts after determination of the cell with the highest score. The direction is now other way around, to upper left, up and left. The scores in these locations are checked and the highest value is selected. There may be more than on highest values which causes alternative alignments. For example, in Figure 2.12 there are two alternatives from the cell with 0.49: *up* and *upper left* since they have the same value. This back track process continues until zero.



G:	0,0	0,0	0,0	0,0	0,3	0,4	0,0	0,0	0,0	0,0	0,0	0,0
A:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
S:	0,0	0,2	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
P:	0,0	0,4	0,2	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
V:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
T:	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
C:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
I:	0,0	0,0	0,2	0,3	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0
N:	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
D:	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Q:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
K:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
E:	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
M:	0,0	0,0	0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0
H:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
F :	0,0	0,0	0,2	0,2	0,2	0,3	0,0	0,0	0,0	0,0	0,0	0,0
R:	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Y:	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0

	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
S	0.0	0.0	0.18	0.11	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	0.0	0.0	0.39	0.39	0.18	0.11	0.0	0.0	0.0	0.0	0.0	0.0	0.0
V	0.0	0.0	0.08	0.42	0.42	0.18	0.11	0.0	0.0	0.0	0.0	0.0	0.0
T	0.0	0.0	0.0	0.08	0.49	0.48	0.23	0.11	0.0	0.0	0.0	0.0	0.0

Figure 2.12. Back track. From the cell with the highest score, three cells are checked in order to find the biggest scores between them. While back track, there may be more than one possible cell with the same highest score (Here, the second cell can be $M[3][4]$ and [4][4]. It leads to obtain alternative alignments.

After back track, alignments can be obtained. There are three options which are:

1. *Up:* There is a gap on a sequence and “-“ is introduced into sequence part..
2. *Left:* There is a gap on a profile and “-“ is introduced into profile part.
3. *Diagonal:* There is a match and characters from sequence and a profile are introduced.

In Figure 2.12, the best alignment is:

Profile: S P V T
| | |
Sequence: S - V T

Some parameters for profile alignment are listed in Table 2.4.

Table 2.4. Some parameters for profile alignment

Name	**Default**	**Information**
Match score	1	To score when an amino acid matches to itself on a sequence of database
Mismatch score	0	To score when an amino acid does not match to itself on a sequence of database
Gap score at sequence	-1	To score a matrix from left (sequence direction)
Gap score at profile	-5	To score a matrix from up (profile direction).
Substitution matrix (SM)	Optional	It is possible to give a SM file in addition to using a simple or any known SMs.
Use mass constrain	Optional	To select alignments use mass constraint between predicted sequence and precursor mass

2.4. Performance Evaluation

2.4.1. Scoring Function

To determine a scoring function for RAy, different scores based on various ion types are compared. First of all, one sequence is arbitrary selected and additional 100 sequences are randomly generated (Length from 4 to 6 and edit distance scores up to 6, edit distance is a metric used to find out similarity between peptide sequences which calculates the minimum number of insertion, deletions and substitutions between sequences (Ristad, Yianilos, & Member, 1998)). After getting preliminary results, another dataset is constructed which contains 21 sequences with edit distance scores 0-2 to select ion types for scoring functions. To analyze these results, edit distance score is normalized and adjusted to masses of sequences (Equation 2.8). As a result of this analysis, MSPepScorer named scoring function is determined (range: [0-1]) (See Section 3.3.1. Scoring Function Analysis for graphics) (Equation 2.9).

$$MASSS = 0.5 * \left(1 - \frac{|mass_a - mass_b\ |}{\max(mass_a, mass_b)}\right) + 0.5 * \left(1 - \frac{ED}{\max(len_a, len_b)}\right) \qquad (2.8)$$

$$MSPepScore = \frac{CosineSimilarity_{b,y}}{2} + \frac{\left(SharedPeakAbuRatio_{a,b,y}\right)}{2} \qquad (2.9)$$

2.4.2. Parameter Settings

2.4.2.1. Workflow

The workflow principle is illustrated on Figure 2.13. An experimental MS/MS spectrum specific sequence profile is constructed depending on given parameters (See Table 2.3). Then, profile is aligned to sequences in database via Smith-Waterman algorithms with user defined parameters (See Table 2.4). Finally, the outcomes are given with alignment results and SW (Equation 2.10) and MSPepScore scores (Equation 2.9).

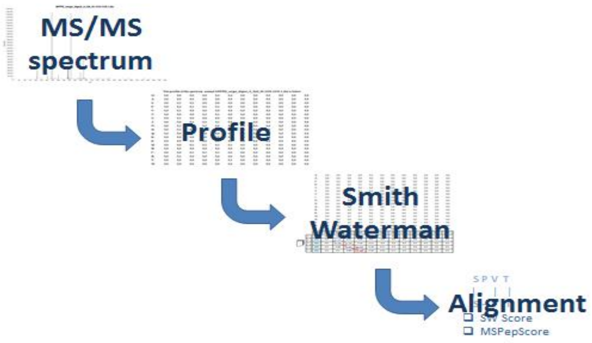


Figure 2.13. The workflow principle.

2.4.2.2. Synthetic Spectra Dataset

Combinations of the scoring system including match, mismatch and gap scores are evaluated in order to get better results from the datasets. Firstly, a mass constrain which shows the mass difference mass of sequence and a precursor mass is used to select alignments. Various mass constrain values are tested (2Da, 5 Da, 10Da, 20Da). Then, Smith-Waterman (SW) scores for alignments are analyzed to set a threshold. The SW score is sum of all frequencies in profile if there is a match while scoring the matrix before alignment (Equation 2.9). After that, the alignments with SW score above the threshold are selected for MSPepScorer analysis for prediction quality.

$$SW_{score} = Sum(frequency_{match}) \tag{2.10}$$

2.4.2.3. Synthetic Peptide and Keller *et al.* Datasets

The best settings are determined from the analysis of the synthetic spectra (See 3.3.2.1. Synthetic Spectra Dataset) are used to evaluate synthetic peptide and Keller *et al.* datasets. Once again, sequence similarity (MASSS) and prediction quality (MSPepScore) are analyzed to see the performance.

CHAPTER 3

RESULTS AND DISCUSSION

3.1. Spectral Quality

Spectral quality of each spectrum in three dataset is assessed to evaluate performance To calculate spectral quality score, ion existence of *a*-, *b*- and *y*-ions on a MS/MS spectraum is checked with +/- 0.3 Da mass tolerance. Then, weighted sum of three ion sets are calculated (Equation 3.1).

The first synthetic peptide and Keller *et al.* datasets (See section 2.1. Tandem Mass Spectra Datasets) are analyzed in respect to spectral quality. For this purpose, ion quality (Ion_Qual) scores are calculated per spectrum (Equation 3.1) (Figure 3.1).

$$Ion\_Qual = (0{,}2 * Num_{a\ ions}) + \left(0{,}4 * Num_{y\ ions}\right) + (0{,}4 * Num_{b\ ions}) \qquad (3.1)$$

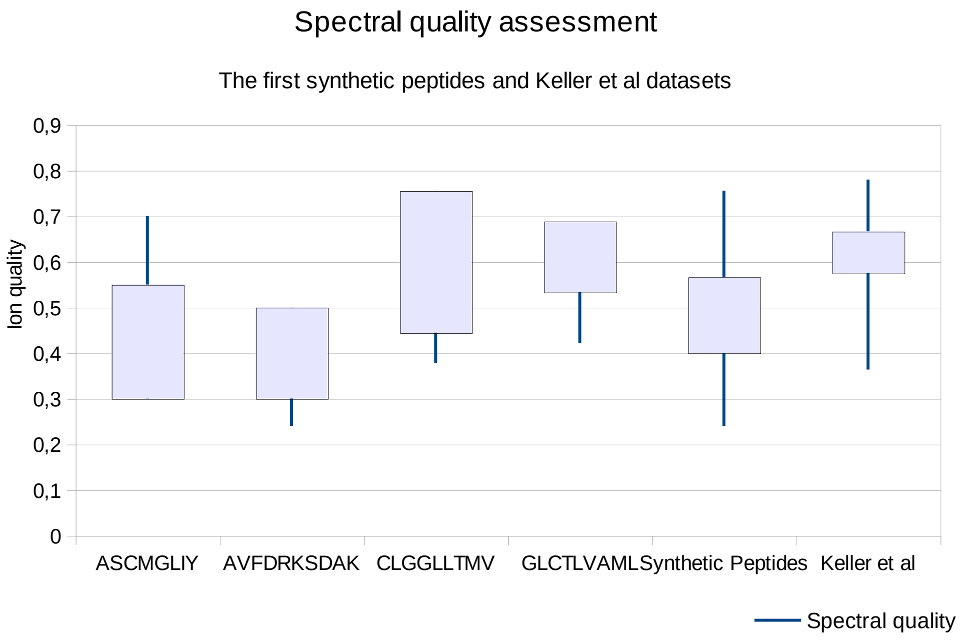


Figure 3.1: Spectral quality assessment of datasets. x-axis shows peptide sequence or dataset name whereas y-axis shows spectral quality. Then weighted sum of found known ion types are calculated for each spectrum. Quality in the first synthetic peptide and Keller *et al.* datasets varies from 0.24 to 0.7

3.2. Profile Construction

3.2.1. Preprocessing Success

The first step in profile construction is spectra preprocessing (See Section 2.3.1.1 Preprocessing Experimental Spectra). Basically, isotopic envelopes are condensed into single peaks. Then, deconvolution of charge state and windows based peak filtration steps are performed.

To evaluate the performance of preprocess steps, existence of found *b*- and *y*-ions are checked in the Keller *et al* dataset. Theoretical spectra with *b*- and *y*-ions depending on expected sequences are constructed. Then, they are compared to experimental spectra regarding to found *b*- and *y*-ions (Equation 3.2). As can be seen in Figure3.2, the majority of the informative peaks are remained.

$$Found\ ion\ ratio = \frac{Number(Found(b_{ions} + \ y_{ions}))}{Theoretical(Found(b_{ions} + \ y_{ions}))} \tag{3.2}$$

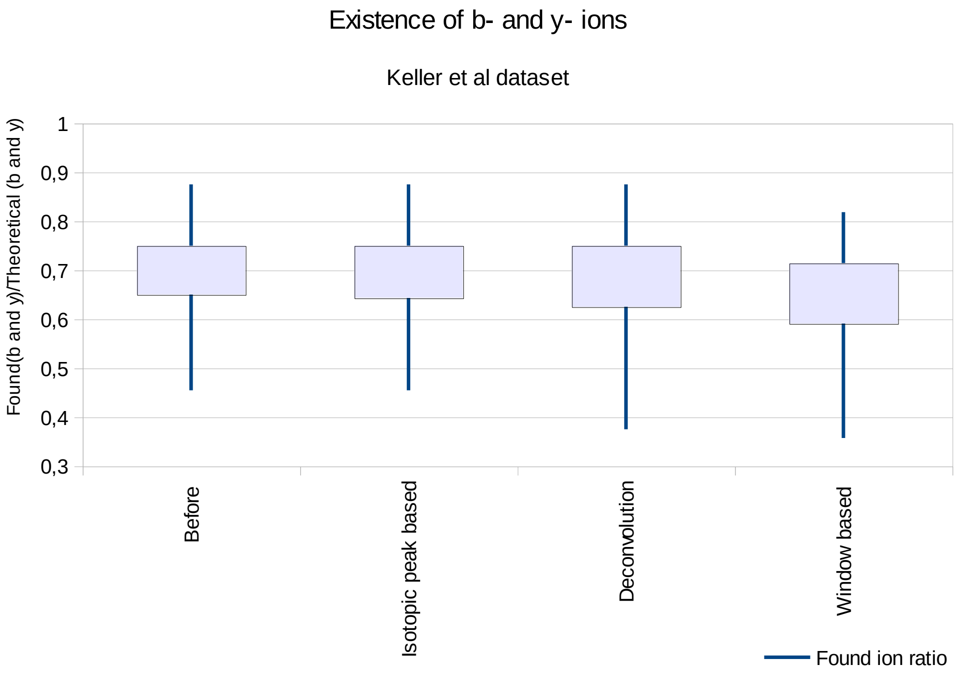


Figure 3.2. Existence status of *b*- and *y*-ions before and after each steps in preprocess which are collapse isotopic peak filtration, deconvolution and windows based peak filtration, respectively. Found ion ratio is calculated based on found theoretical peaks. At the end of preprocessing, majority of *b*- and *y*- ions still exist.

To see the status of found ion ratios over all peaks after preprocessing steps, calculated found ion ratios are normalized with total ion number on a spectrum. To achieve this, found ion ratios are divided by sum of all peaks (Equation 3.3). Figure3.3 shows the distribution of normalized found ion ratios in Keller *et al.* dataset after each step. While peaks are discarding, a remarkable percent of informative peaks still remain.

$$Normalized\ found\ ion\ ratio = \frac{\dfrac{Number(Found(b_{ions} + y_{ions}))}{Theoretical(Found(b_{ions} + y_{ions}))}}{Number(Total\ Peaks)} \qquad (3.3)$$

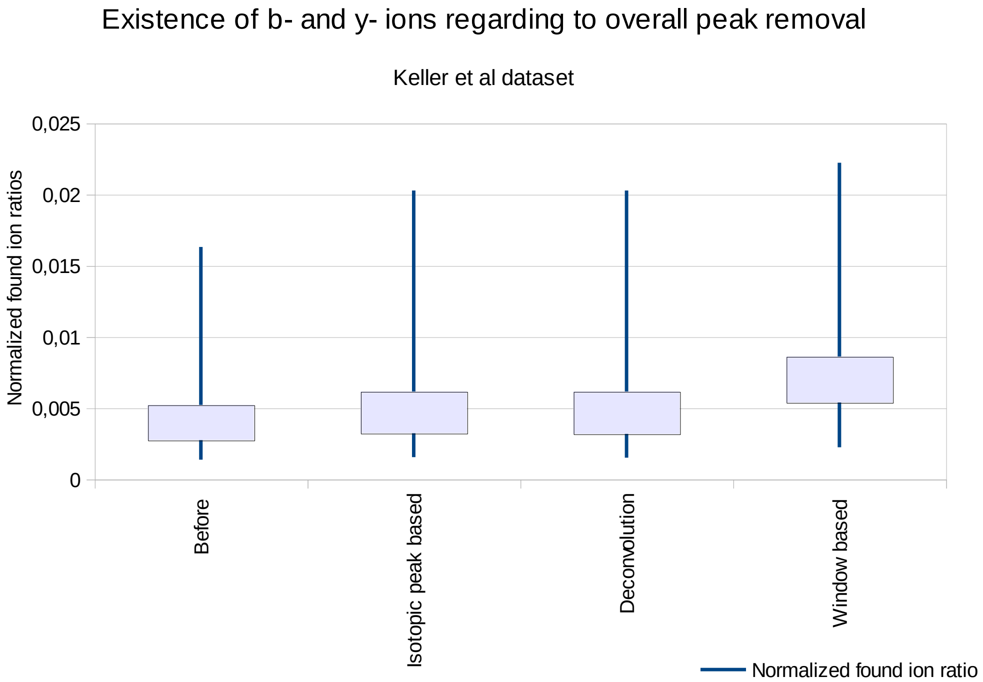


Figure 3.3. Normalized found ion ratios before and after each step in preprocess which are collapse isotopic peak filtration, deconvolution and windows based peak filtration, respectively. Found ion ratios are normalized with total peak number. While peaks are removed, majority of *b*- and *y*- ions still exist.

To have a closer look at overall peak removal success, before and after existences of *b*- and *y*-ions are also checked against overall removed ratios. Retained peaks number ratio (RPR) is the proportion of number of *b*- and *y*-ions that are found before and after preprocess steps (Equation 3.4). Overall removed peak number ratio (ORPR) is also the proportion of total removed peak number and the value that total peaks is subtracted by found *b*- and *y*-ions after process (Equation 3.5).

$$RPR = \frac{After(total(b\ ions + y\ ions))}{Before(total\ (b\ ions + y\ ions))} \qquad (3.4)$$

$$ORPR = \frac{Num(total(Removed\ Peaks))}{Before\big(Num(total\ Peaks)\big) - After\big(Num(bions + y\ ions)\big)} \qquad (3.5)$$

Then, retained peak ratio values (RPR) are plotted againes overall removed peak ratio (ORPR) in Figure 3.4. It is clear that eventhough peak removal, informative peaks which are *b*- and *y*-ions still exist. Moreover, due to deconvulation, some doubly charged peaks are successfully converted to singly charged status. Therefore, the retained peak ratio is bigger than 1.0 in some cases.

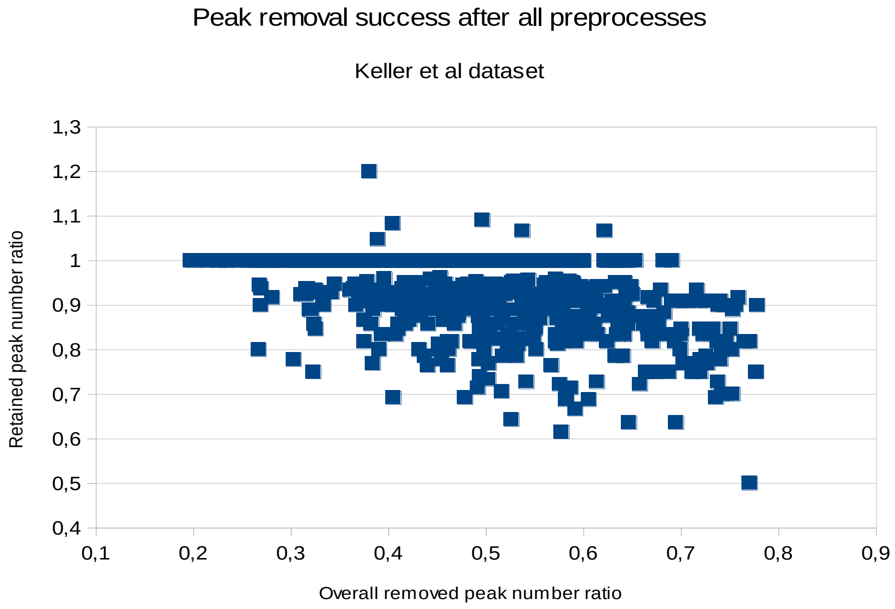


Figure 3.4. Peak removal success. *x*-axis shows overall removed peak ratio(ORPR) whereas *y*-axis indicated retained peak number ratio (RPR) values. After preprocessing, many *b*- and *y*-ions still exist. Besides, due to deconvolution of doubly charged peak into singly charge status, there are some values that are bigger than 1.0 in *y*-axis.

In addition, doubly charged peak removal success is also analyzed in Keller *et al.* dataset. In this step, 656 spectra with doubly charged precursor ion and also 18 spectra with triply charged precursor ion in the dataset are analyzed. Removed doubly charged peak ratio (RDCR) is calculated (Equation 3.5) and then ratios are analyzed in respect to spectral quality groups (Figure 3.5). The Figure 3.5 reveals that the majority of doubly charged peaks are removed after process. Furthermore, there is a correlation between spectral quality and RDCR especially for the spectra with triply charged precursor ions.

$$RDCR = \frac{After\big(Total(b^{2+} + y^{2+})\big) - Before(Total(b^{2+} + y^{2+})}{Before(Total(b^{2+} + y^{2+})} \qquad (3.5)$$

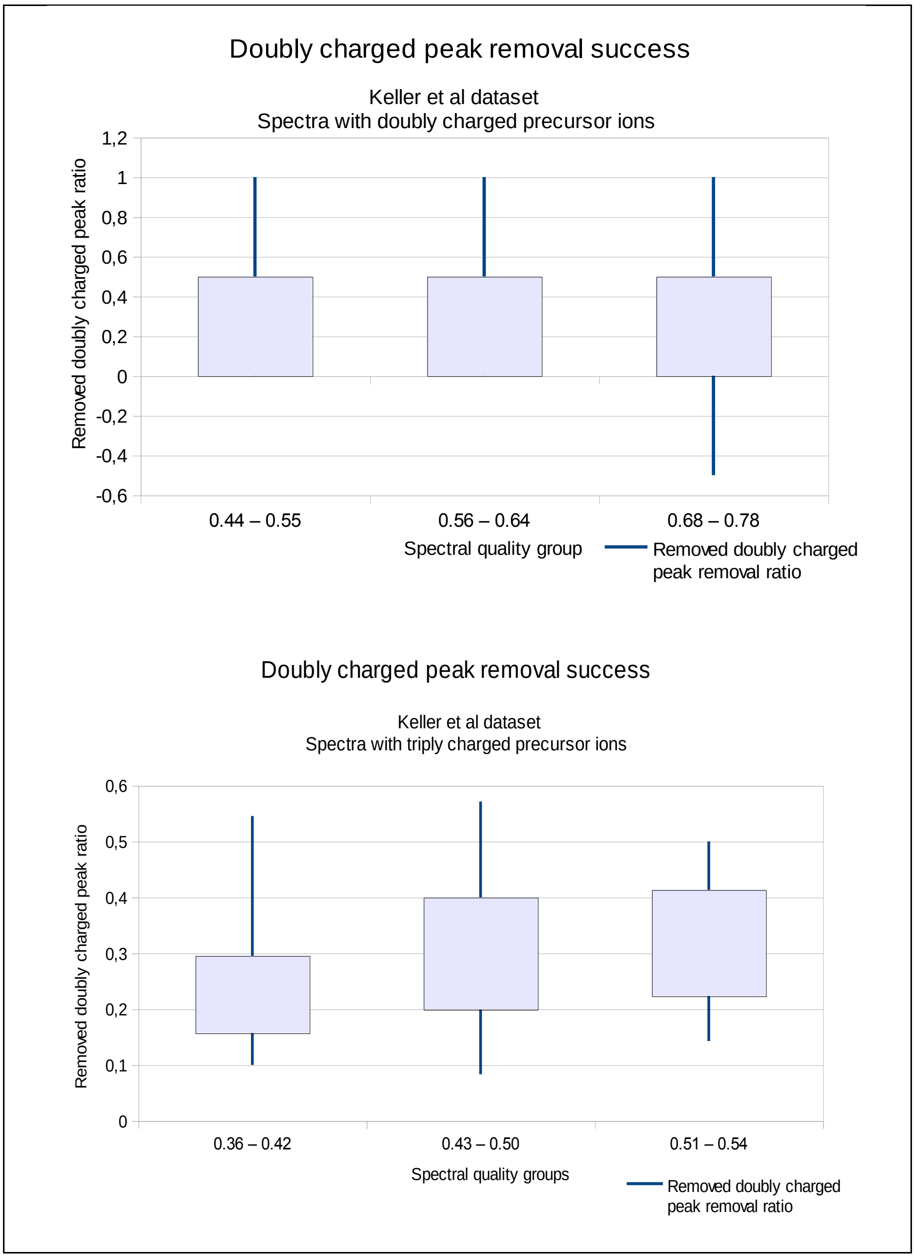


Figure 3.5. Doubly charged peak removal success. x-axis shows spectral quality groups whereas y-axis shows removed doubly charged peak ratio (RDCR) distribution. 629 spectra with doubly charged precursor ion and 18 spectra with triply charged precursor ion in Keller *et al.* dataset are analyzed.

3.2.2. Orientation

One of the essential features on RAy is orientation elimination as much as possible. Herein, two aspects are used (See Section 2.3.1.3. Orientation Elimination): the first strategy leads increase score of individual fit amino acids in orientation eliminated way and the second approach is regarding to correction of directionalities in tags.

3.2.2.1. Fit Amino Acids

To evaluate the success of orientation elimination on individual fit amino acids, Keller *et al.* dataset is used. Since the expected sequences in each experimental spectrum is known, it is checked whether fit amino acid is a part of the sequences, and then location is determined (right or left part of expected sequences). So as to calculate orientation score, all fit amino acids with the same direction are counted and then divided by the overall fit amino acid sequences. This process is repeated after orientation elimination step. If the orientation score equals to 1, it shows that all fit amino acids are in the same direction. If the scores in both directions equals to each other, orientation score is 0. The analysis is plotted in Figure 3.6. After this process in orientation elimination part, orientation scores and also amino acids with the same directions are increased. There is a correlation between spectral quality and orientation elimination success, orientation elimination works better on the spectra with high quality.

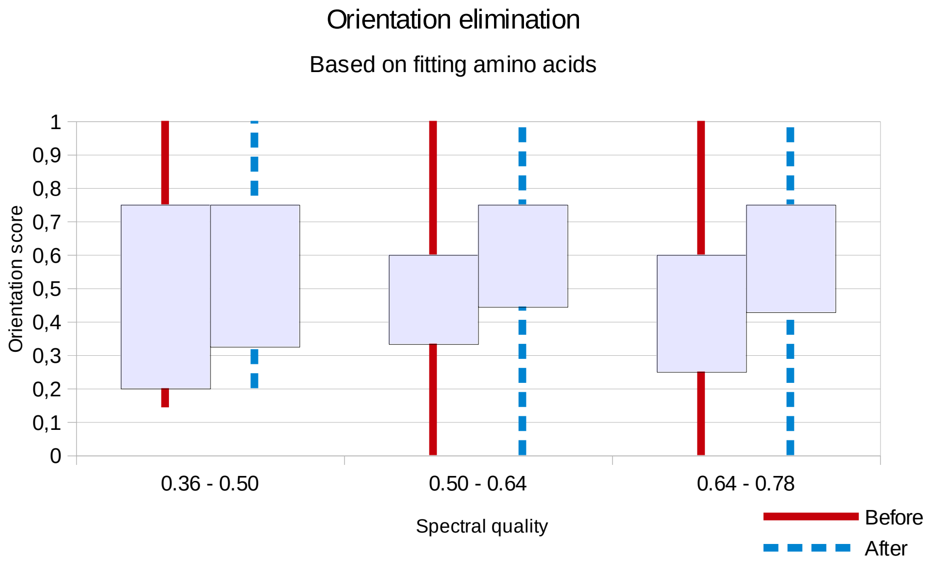


Figure 3.6. Orientation elimination, based on fit amino acids. *x*-axis shows spectral quality group and *y*-axis indicates orientation score distributions. Solid and dashed lines next to each other show the maximum and minimum values before and after orientation, respectively. Status756 spectra in Keller *et al.* dataset are used in this analysis, regardless of the charge status of precursor ions. Orientation scores for before and after the process are calculated and then compared. While spectral quality is increasing, orientation elimination success is improved as well.

3.2.2.2. Tags

To evaluate the performance of tag orientation elimination process (See Section 2.3.1.3. O, tags of each spectrum on the Keller *et al.* and synthetic spectra datasets are analyzed. After preprocessing of each spectrum, tag directions are found out since the expected sequence of that spectrum is known. If amino acids in tag sequence are mainly in the direction from the left to the right, they are considered “forward”; otherwise they are “reverse”. For an example in Figure 3.7, the tag with sequence “GASP” is expected to be left since amino acids are from the left to the right according to the expected sequence which is “GASPVT”. Based on this information, all directions for the tags before and after orientation are named. Then, orientation scores (Equation 3.6) are calculated for each spectrum and the score distributions for the different datasets are plotted in Figure 3.8.

$$Score_{orientation} = \frac{Count(The\ Tags\ with\ the\ most\ frequent\ direction)}{Count(All\ Tag)} \quad (3.6)$$

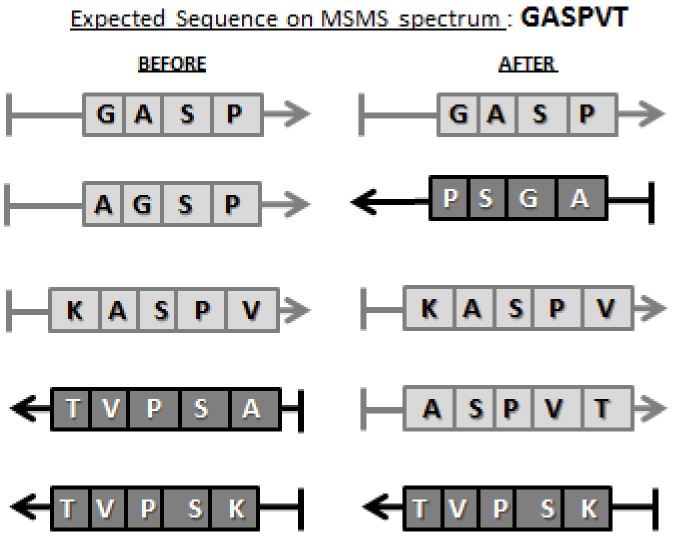


Figure 3.7. Orientation success analysis. Herein, the expected sequence is “GASPVT”. According to this information, the directionalities of found tags are determined whether “forward” (light gray) or “reverse” (dark gray).

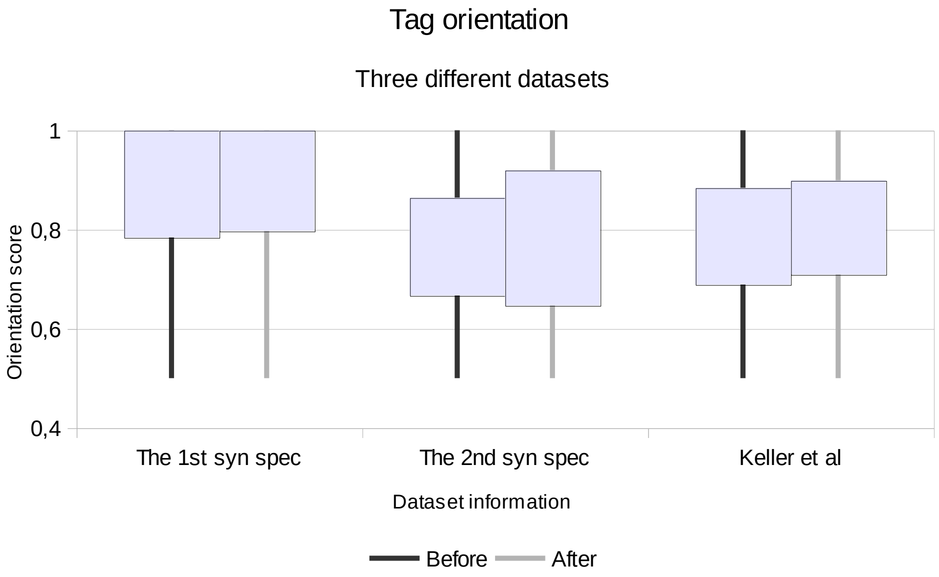


Figure 3.8. Tag orientation process success. x-axis shows dataset information which are the first synthetic spectra and the second synthetic spectra and Keller *et al.* datasets, respectively. On the left of each group shows the orientation score distributions before process (dark gray) and on the right indicates the distributions after (light gray). After orientation elimination process, generally scores are increased for each dataset.

3.3. Profile Alignment

3.3.1. Scoring Function Analysis

In order to determine the scoring function of RAy, the analysis explained on the section “2.4.1 Scoring Function Analysis” is carried out. The dataset with 21 sequences are used to select ion types for the selected scoring functions. To evaluate outcomes, edit distance score is normalized and adjusted to mass of the sequences (called as mass adjusted sequence similarity score - MASSS) (Equation 3.7). This score enables differentiating the sequences which have same edit distance score, but different masses. After evaluation of ion type combinations, the weighted sum of shared peak abundance ratio (a-, b- and y-ions) and cosine similarity (b- and y-ions) scores (Figure 3.9) is decided to be used and this score is called as MSPepScore (Equation 3.8) (range: [0-1]).

$$MASSS = 0.5 * \left(1 - \frac{|mass_a - mass_b\,|}{\max(mass_a, mass_b)}\right) + 0.5 * \left(1 - \frac{ED}{\max(len_a, len_b)}\right) \quad (3.7)$$

$$MSPepScore = \frac{CosineSimilarity_{b,y}}{2} + \frac{\left(SharedPeakAbuRatio_{a,b,y}\right)}{2} \quad (3.8)$$

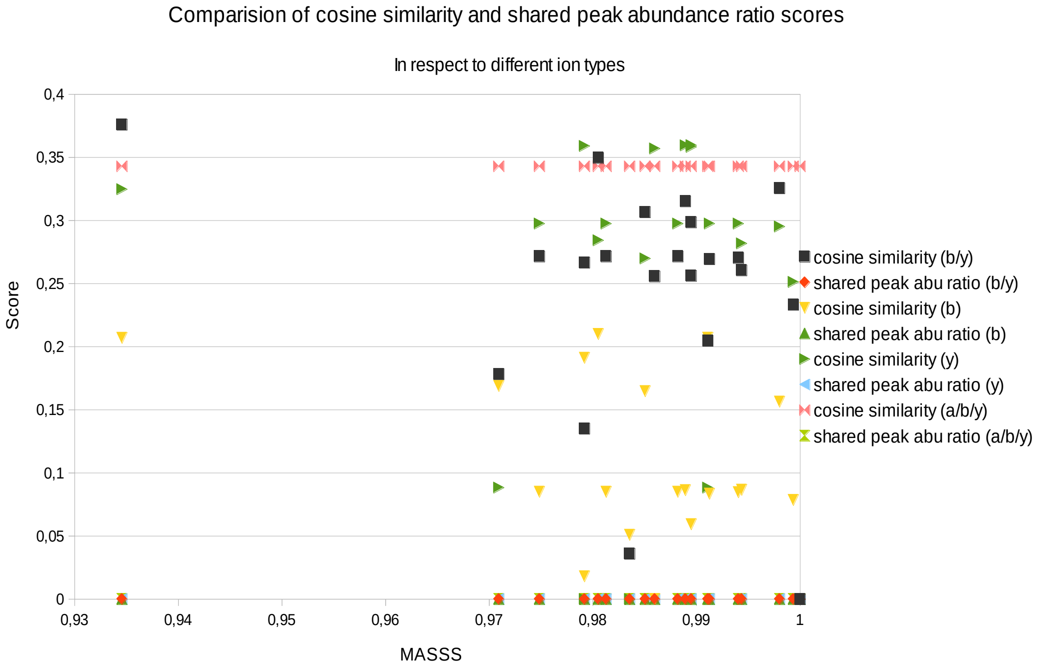


Figure . 3.9. Comparison of cosine similarity and shared peak abundance ratio scores based on different ion types. *x*-axis shows MASSS which is mass adjusted sequence similarity score and *y*-axis indicated different scoring functions. 8 different scores are compared and as a result. The combination of shared peak abundance ratio (*a*-, *b*- and *y*-ions) and cosine similarity (*b*- and *y*-ions) scores (Equation 3.7) is decided to be used as a MSPepScore.

3.3.2. Parameter Settings Evaluation

3.3.2.1. Datasets

The analysis to evaluate alignments the different spectra in the datasets with various parameters are carried out. Herein, sequence similarity (edit distance or MASSS) and scoring function (MSPepScore) are being analyzed. We aim to get the best results for the sequence that exactly matches to expected sequences, better result for the sequences that have some variation (addition/insertion/mutation) of expected sequences and the worst results for the sequences that are completely different than expected sequences.

CHAPTER 4

CONCLUSION

Mass spectrometry (MS) is a commonly used technique in protein studies. There are two main strategies regarding computational aspects to interpret mass spectrometric data: database search and *de novo* sequencing. The success of database search algorithms depends on correctly annotated sequences. In case that sequence is missing, database searches fail. If a sequence of protein in study is absent in database, homology search can be useful.

We are presenting a new algorithm named RAy which aims to increase the scope of MS based homology searches in sequence databases. It has a novel approach – sequence *profile* which is specific for each MS/MS spectrum. RAy has two steps: *profile construction* and *profile alignment.* After construction of sequence profile, they are directly aligned to sequences in a database via the Smith-Waterman algorithm.

The first steps of RAy in profile construction are preprocessing including isotopic peak filtration, deconvolution. After deconvolution, the number of the informative ions is increased. Also, filtration steps are successful since *b*- and *y*-ions are still remaining while removing unknown ions. RAy tries to eliminate directionality depending on the two strategies which are fit amino acid and tag based. Generally algorithm in computational mass spectrometry work based on tryptic cleavage of peptides, however RAy offers a possibility to work with the spectra which proteins are not cleaving by any tryptic enzymes. Moreover, alignment results are not restricted to mass constrain which is mass difference between sequence at database and precursor ion of a tandem MS spectrum and that is an optional settings can be changed by a user.

REFERENCES

Aebersold, R, & Goodlett, D. R. (2001). Mass spectrometry in proteomics. *Chemical reviews*, *101*(2), 269-95.

Aebersold, Ruedi, & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature, 422*(6928), 198-207.

Allmer, J. (2011). Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Review of Proteomics*, *8*(5), 645-657.

Bandeira, N., Olsen, J. V., Mann, M., Mann, M. & Pevzner, P. A. (2008). Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. *Bioinformatics*, *24*(13), i416-i423.

Cao, X., & Nesvizhskii, A. I. (2008). Improved Sequence Tag Generation Method for Peptide Identification in Tandem Mass Spectrometry, American Chemical Society, 7(10), 4422-4434.

Chait, B. T. (2006). Mass spectrometry: bottom-up or top-down? *Science, 314*(5796), 65-6.

Cottrell, J. S. (2011). Protein identification using MS/MS data. *Journal of Proteomics, 74*(10), 1842-51.

Cox, J., & Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. *Annual Review of Biochemistry*, *80*, 273-99.

Craig, R., & Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics, 20*(9), 1466-1467.

Dancík, V., Addona, T. A, Clauser, K. R., Vath, J. E., & Pevzner, P. A. (1999). De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *6*(3-4), 327-42.

Domon, B., & Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science, 312*(5771), 212-7.

Edman, P. (1950). Method for determination of the amino acid sequence in peptides. *Acta Chemica Scandinavica*, 283-93.

Eidhammer, I., Flikka, K., Martens, L., & Mikalsen, S.O. (2008). *Computational Methods for Mass Spectrometry, Chichester, UK, John Wiley & Sons, Ltd (1st Ed.,* p. 296).

Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, *Nature Methods, 4*(3), 207-214.

Elias, J. E., & Gygi, S. P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics, *Methods Mol Biol., 604*, 55-71.

Eng, J. K., Mccormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Sciences, 67*(8), 1426-36.

Forner, F., Foster, L. J., & Toppo, S. (2007). Mass spectrometry data analysis in the Proteomics era. *Current Bioinformatics, 2*(1), 63-93.

Frank, A., & Pevzner, P. (2005). PepNovo De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry, 77*(4), 964-973.

Frank, A., Tanner, S., Bafna, V., & Pevzner, P. (2005). Peptide Sequence Tags for Fast Database Search in Mass-Spectrometry. *American Chemical Society, 4(*4), 1287 – 1295.

Geer, L. Y., Markey, S. P., Kowalak, J. A, Wagner, L., Xu, M., Maynard, D. M., Yang, X., et al. (2004). Open mass spectrometry search algorithm. *Journal of Proteome Research, 3*(5), 958-64.

Gentzel, M., Köcher, T., Ponnusamy, S., & Wilm, M. (2003). Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics, 3*(8), 1597-610.

Gevaert, K., & Vandekerckhove, J. (2000). Protein identification methods in Proteomics. *Review Electrophoresis, 21*, 1145-54.

Han, Y., Ma, B., Zhang, K., & Na, C. (2004). SPIDER : Software for Protein Identification from Sequence Tags with De Novo Sequencing Error. *Proc IEEE Computational Systems Bioinformatics Conference*, 206-15.

Hellerstein, M. K., Neese, R. A., (1999). Mass isotopomer distribution analysis at eight years : theoretical , analytic , and experimental considerations Mass isotopomer distribution analysis at eight years : theoretical , analytic , and experimental considerations. *Am J Physiol Endocrinol Metab*, 1146-1170.

Henzel, W. J., Watanabe, C., & Stults, J. T. (2003). Protein identification: The origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry, 14*(9), 931-42.

Huang, L., Jacob, R. J., Pegg, S. C., Baldwin, M. a, Wang, C. C., Burlingame, A. L., & Babbitt, P. C. (2001). Functional assignment of the 20 S proteasome from *Trypanosoma brucei* using mass spectrometry and new bioinformatics approaches. *The Journal of Biological Chemistry, 276*(30), 28327-39.

Johnson, R. S., Davis, M. T., Taylor, J. A., & Patterson, S. D. (2005). Informatics for protein identification by mass spectrometry. *Methods*, *35*(3), 223-36.

Kelleher, N. L. (2004). Top-down proteomics. *Analytical chemistry*, *76*(11), 197-203.

Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., & Kolker, E. (2002). Experimental protein mixture for validating tandem mass spectral analysis. *Omics*, *6*(2), 207-212.

Kim, S., Bandeira, N., & Pevzner, P. A. (2009). Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. *Molecular & Cellular Proteomics: MCP, 8*(6), 1391-400

Kiner, M., & Sherman, N. E. (2000). *Protein sequencing and identification using tandem mass spectrometry* (1st ed.). Wiley-Interscience.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., & Lajoie, G. (2003). PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry : RCM*, *17*(20), 2337-42.

Ham, A. J., Aerni, H. R., Cheek, K., Whitwell, C. W., Caprioli, R. M. , Tabb D. L. , Ma, Z., (2011). ScanRanker: Quality assessment of tandem mass spectra via sequence tagging. *Journal of proteome research*, *10*(7),

Mackey, A. J., Haystead, T. A. J., & Pearson, W. R. (2001). Getting More from Less: Algorithms for Rapid Protein Identification with Multiple Short Peptide Sequences. *Molecular & Cellular Proteomics:MCP*, *1*(2), 139-147.

Mann, M, & Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, *66*(24), 4390-9.

Mann, Matthias, Hendrickson, R. C., & Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annual Review of Biochemistry*, *70*, 437-73.

Matthiesen, R. (2006). *Mass Spectrometry Data Analysis in Proteomics* (1st ed., p. 336). Humana Press.

McHugh, L., & Arthur, J. W. (2008). Computational methods for protein identification from mass spectrometry data. *PLoS computational biology*, *4*(2), e12.

Olsen, J. V., & Mann, M. (2004). Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(37), 13417-22.

Perkins, D. N., Pappin, D. J. . C., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data Proteomics and 2-DE. *Electrophoresis*, *20*(18), 3551-3567.

Ristad, E. S., Yianilos, P. N., & Member, S. (1998). Learning String-Edit Distance, *IEEE Transactions on Pattern Recognition and Machine Intelligence* , *20*(5), 522-532.

Searle, B. C., Dasari, S., Turner, M., Reddy, A. P., Choi, D., Wilmarth, P. A., Mccormack, A. L., David, L.L., Nagalla, S.R. (2004). High-Throughput Identification of Proteins and Unanticipated Sequence Modifications Using a Mass-Based Alignment Algorithm for MS / MS de Novo Sequencing Results. *Analytical Chemistry*, *76*(8), 2220-2230.

Seidler, J., Zinn, N., Boehm, M. E., & Lehmann, W. D. (2010). De novo sequencing of peptides by MS/MS. *Proteomics*, *10*(4), 634-49. doi:10.1002/pmic.200900459

Shadforth, I., Crowther, D., & Bessant, C. (2005). Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. *Proteomics*, *5*(16), 4082-95.

Shevchenko, A., Sunyaev, S., Loboda, A., Bork, P., Ens, W., & Standing, K. G. (2001). Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Analytical Chemistry*, *73*(9), 1917-26

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195-197.

Standing, K. G. (2003). Peptide and protein de novo sequencing by mass spectrometry. *Current Opinion in Structural Biology*, *13*(5), 595-601.

Steen, H., & Mann, M. (2004). The ABC’s (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, *Molecular cell biology*, *5*(9), 699-711.

Sun, H., Zhang, J., & Liu, H. (2011). ., A New Scoring Scheme for Peptide Sequence Tagging via Doubly Charged MS / MS Spectra, *Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference on*, Wuhan China 1-4.

Sunyaev, S., Liska, A. J., Golod, A., Shevchenko, A., & Shevchenko, A. (2003). MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Analytical chemistry*, *75*(6), 1307-15.

Sykes, M. T., & Williamson, J. R. (2008). Envelope: Interactive software for modeling and fitting complex isotope distributions. *BMC Bioinformatics*, *9*(1), 446.

Tabb, D. L., Fernando, C. G., & Chambers, M. C. (2007). MyriMatch : Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis, *Journal of Proteome Research*, 6(2), 654-661.

Tabb, D. L., Ma, Z. Q., Martin, D. B., Ham, A.J., & Chambers, M. C. (2008). DirecTag : Accurate Sequence Tags from Peptide MS / MS through Statistical Scoring research articles. *Journal of Proteome Research*, *7(*9), 3838-3846.

Tabb, D. L., Saraf, A., & Yates, J. R. (2003). GutenTag : High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. *Analytical Chemistry*, *75*(23), 6415-6421.

Tanner, S., Shu, H., Frank, A., Wang, L.-chi, Zandi, E., Mumby, M., Pevzner, P. A., Bafna, V., (2005). InsPecT : Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra. *Analytical Chemistry*, *77*(14), 4626-4639.

Taylor, J. A, & Johnson, R. S. (1997). Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry : RCM*, *11*(9), 1067-75.

Tuli, L., & Ressom, H. W. (2009). LC-MS Based Detection of Differential Protein Expression. *Journal of Proteomics & Bioinformatics*, *2*(10), 416-438.

UCSF, Directed by Dr. Alma Burlingame, *MS-Product.* http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct (accessed July 22, 2012).

Uniprot, http://www.uniprot.org/ (accessed July 22, 2012).

Xu, C., & Ma, B. (2006). Software for computational peptide identification from MS-MS data. *Drug Discovery Today*, *11*(13-14), 595-600.