September 2015

Algorithms for Peptide Identification from Mixture Tandem Mass Spectra

Yi Liu
The University of Western Ontario

Supervisor
Kaizhong Zhang
The University of Western Ontario

Graduate Program in Computer Science

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

ALGORITHMS FOR PEPTIDE IDENTIFICATION FROM MIXTURE TANDEM MASS SPECTRA

(Thesis format: Monograph)

by

Yi Liu

Graduate Program in Computer Science

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

The large amount of data collected in an mass spectrometry experiment requires effective computational approaches for the automated analysis of those data. Though extensive research has been conducted for such purpose by the proteomics community, there are remaining challenges, among which, one particular challenge is that the identification rate of the MS/MS spectra collected is rather low. One significant reason that contributes to this situation is the frequently observed mixture spectra, which result from the concurrent fragmentation of multiple precursors in a single MS/MS spectrum. The ability to efficiently and confidently identify mixture spectra is essential to alleviate the current bottleneck of low mass spectra identification rate. However, nearly all the mainstream computational methods still take the assumption that the acquired spectra come from a single precursor, thus they are not suitable for the identification of mixture spectra.

In this research, a mixture spectrum is modelled as linear combination of two single-precursor spectra and we focused on developing effective algorithms for the purpose of interpreting mixture tandem mass spectra. Our research work is mainly comprised of two components: mixture spectra *de novo* sequencing and mixture spectra identification by database search. For the *de novo* sequencing approach, we formulated the mixture spectra *de novo* sequencing mathematically, and proposed a dynamic programming algorithm for the problem. Different from the conventional idea of seeking multiple peptides from a MS/MS spectrum iteratively, our proposed algorithm considers the problem as a whole and proceeds in a rigorously designated pathway to construct two peptide sequences concurrently from a single mixture spectrum. Additionally, we use both simulated and real mixture spectra datasets to verify the efficiency of the algorithm described in the research. For the database search identification, we proposed an approach for matching mixture tandem mass spectra with a pair of peptide sequences acquired from the protein sequence database by incorporating a special *de*

novo assisted filtration strategy. Prior to scoring the query mixture spectrum against the enormous amount of candidate pairs by a stringent scoring function, the *de novo* assisted filtration strategy will rank all the peptides that are obtained directly from the theoretical digestion of the protein sequence database by an initial filtration scoring model and only those higher ranked peptide sequences are selected and paired up to form candidate pairs that can go through the subsequent matching procedure with a more stringent scoring function. Besides the filtration strategy, we also introduced in the research a method to give an reasonable estimation of the mixture coefficient which represents the relative abundance level of the co-sequenced precursors. The preliminary experimental results demonstrated the efficiency of the integrated filtration strategy and mixture coefficient estimating method in reducing examination space and also verified the effectiveness of the proposed matching scheme.

Keywords: Mass Spectrometry, Computational Proteomics, Mixture Spectra, Peptide *De Novo* Sequencing, Database Search

Acknowlegements

Time flies! It is coming to an end for my doctoral study. It has been a passionate and diligent adventure for me in the past four years. This is a moment to retrospect the past , to seize the present, and to look to the future.

First and foremost, I want to express my sincere gratitude to my doctoral advisor, Dr. Kaizhong Zhang. I still remembered the first time I met Dr. Zhang. It was in Nov. 2010 at Central South University after him participating a research seminar. At that time, I conveyed to him my wish of pursuing the doctoral study under his supervision. Actually I was really nervous talking with him at first, but his gentle words made me calm down. And starting from that moment, I knew he is very easy-going. It was a great pleasure to work under his mentoring in those years afterwards, and I knew that without the support from his forever considerate thought, his solid mathematical and algorithmic knowledge, and his profound research experience, I can't complete the doctoral study. Thank you very much.

Also, I want to thank all my lab colleagues for their accompanying in the past four years. There are so many laughters reflecting in my memory, and so many research ideas sparkling during the discussion. Thank you very much, Dr. Weiming Li, Dr. Lin He, Zhewei Liang, Fang Han, Qin Dong, Yiwei Li, Wenjing Wan, and Yu Qian.

Moreover, I want to thank all our collaborators. Dr. Bin Ma, Dr. Gilles Lajoie, for their insightful suggestions when discussing with them. And I also want to thank Dr. Lucian Ilie and Dr. Lila Kari for spending their time on reading my Topic/Survey Proposal and providing useful advices for my research.

Specially, I am sincerely grateful to my girlfriend Weiping Sun, the soon-to-become doctor, for her always being nice and thoughtful. Every day is a wonderful day being with you and

I can't imagine what a mess my life will be without you being around. Also, I want to thank my pet cat, Xiaohua. She is a lovely cat, but it always bothers me when dealing with her litter box.

Lastly, I want to thank my parents, there will never be enough words to express my gratitude to you. And at this moment, perhaps, no-words is better than any-words.

Thank you all!

I always believe that an end to one thing is indeed the beginning of another. I know there will be new challenges waiting for me in the future. New challenge also means new opportunity, and I am ready to face the incoming adventure with passion and courage.

To my dear family, for their always being supportive during my doctoral adventure

Contents

List of Figures

List of Tables

List of Appendices

List of Algorithms

Chapter 1

Introduction

1.1 Background

Proteins are essential to life. They play key roles in all biological processes. The backbone structure of a protein is a sequence of amino acids; and there are 20 different amino acids commonly observed in living things. Because of the significance of proteins for living beings, proteomics, the subject that systematically studies proteins has gradually become fundamental in the research related to molecular biology. Its primary objective is to obtain a comprehensive understanding of disease formation, cellular processes and interaction activities at the protein level [1]. Efficient analysis of proteomics data will be beneficial for the discovery of biomarkers, which will be extremely useful in the diagnostic and therapeutic procedure of several kinds of diseases in modern clinical treatment and basic medical research [2]. Nowadays, the proteomics research highly depends on the successful identification and quantification of proteins that are expressed in a specific cell, tissue or organism.

During the past two decades, mass spectrometry has gradually become a standard technique for the high-throughput characterization of large biomolecules, including peptides and proteins [3, 4]. In a typical LC-MS/MS experiment, protein sample is digested into peptides

with proteolytic enzymes to break protein molecules into relatively short peptide sequences, and the resulting peptide mixtures are separated using Liquid Chromatography (LC) first, and subsequently ionized using Matrix-Assisted Laser Desorption Ionization (MALDI) [5] or Electrospray Ionization (ESI) [6]. After ionization, the charged peptides are measured in the first mass analyzer, and then fragmented and followed by a measurement in the second mass analyzer. Usually two types of spectra are collected, MS spectra (or survey scans) in which the intensity and m/z are measured for intact peptides and MS/MS spectra where one of the ions detected in the MS spectra is isolated, fragmented and measured in a high-throughput manner [7]. Equipped with high sensitivity and accuracy, current mass spectrometers can produce thousands of MS/MS spectra in a single run. The large amount of data collected in an MS experiment requires effective computational approaches to automate the process of spectra interpretation.

Currently, much efforts have been made to the development of approaches for the computational analysis of mass spectrometry based proteomics data. Generally, the mainstream computational methods for this purpose fall into two categories: database search and *de novo* sequencing. The database search method has been extensively studied, in which the identification of MS/MS spectra is assisted with a protein sequence database, and the primary task is to correctly correlate the collected spectra with some amino acid sequences in the protein database. Many software packages are available for this purpose, including MOWSE [8], Mascot [9], PEAKS DB [10], SEQUEST [11], X!Tandem [13], OMSSA [12] and Phenyx [14]. Usually, methods taking this approach make the assumption that all the sequences in the database are accurate and the proteins in the sample are included in the database. However, the aforementioned prerequisite that the targeted sequence is contained in the database is often not satisfied due to many reasons, such as incomplete genome sequencing, inferior gene prediction from the genome, and the existence of mutations and polymorphisms in the sample. Under this circumstance, *de novo* sequencing will serve as a complementary approach for peptide identi-

fication. In the *de novo* sequencing, the computation of peptide sequence does not rely on the protein database, the algorithm directly constructs the peptide sequence that best matches the spectra. In the earlier days, researchers used to apply a very simple algorithmic approach to deal with peptide *de novo* sequencing by enumerating every possible amino acid combinations for a given molecular mass value, this will result in the exponential increase in computation time with respect to the sequence length, which makes it infeasible for long peptide sequence. In the literature [15], a new mathematical strategy for *de novo* sequencing was introduced. In this research, an MS/MS spectrum is converted to the corresponding spectrum graph and the searching for solution is primarily done on the graph. After this publication, other researchers have proposed various methods for *de novo* sequencing based on the graph model and its variations. For the purpose of *de novo* sequencing, there are also several software packages, including Lutefisk [16], PEAKS [17], PepNovo [18], pNovo [19], EigenMS [20], Sherenga [21], NovoHMM [22], MSNovo [23], AUDENS [24], SEQUIT [25], PPM-Chain [26]. Other reports regarding the *de novo* sequencing can be found in [27, 28, 29, 30, 31, 32]. There are also literature on the review and comparison of the commonly accepted *de novo* sequencing methods available to us, in which the research will provide us more comprehensive understanding of the *de novo* sequencing approach and the related algorithms [33, 34, 35, 36, 37].

Although much effort has been made to develop new computational approaches for the analysis of mass spectrometry data, there are still several unsolved problem that are challenging [38]. One specific challenge is that in a high throughput MS/MS experiment, usually only a fraction of the acquired spectra can be confidently interpreted by the existent computational methods. Many factors may contribute to this situation which include: low precursor intensity, poor fragmentation of the selected precursor, or the existence of modified residues. Moreover, the gas phase fragmentation may result in MS/MS spectra with unconventional fragment ions that are not considered by the mainstream computational methods [39], and the sequenced peptides may not be present in the database or may have unanticipated post-translational mod-

ifications (PTMs) [40]. Another specific scenario that facilitates the low identification problem is caused by the ion isolation process in the LC-MS/MS experiment. Recall that the precursor ion is selected in the first mass analyzer in tandem mass spectrometer based on its m/z value. Even coupled with *High Performance Liquid Chromatography (HPLC)* [41], it is still not guaranteed that the peptides in the sample are completely separated. There is a chance that peptides with similar m/z values co-elute, generating a single spectrum that contains a mixture of spectra. The mixture spectra are induced by the isolation and simultaneous fragmentation of two or more distinct molecular ions within the same isolation window. Fragments from multiple precursors will be present in a single MS/MS spectrum, increasing the number of unidentified fragments in database search engines.

A considerable portion of the spectra acquired in the typical *Data Dependent Acquisition (DDA)* strategy [42] come from the concurrent fragmentation of multiple precursors in the same sequencing attempt. Previous research has confirmed that peptides with similar mass values and chromatographic properties being sequenced together can happen quite frequently and result in mixture spectra containing ion fragments from multiple precursors [43, 44, 45, 46]. In [43], Hoopmann and Finney examined the frequency of mixture spectra with a software tool and estimated that 11 percent of MS/MS spectra are chimeras, with an additional 29 percent of MS/MS whose parent isotope distribution inconsistent with peptide analytes. In [44], Houel et al. investigated the frequency of *chimeras* in shotgun proteomics and assessed that in a typical data-dependent acquisition (DDA) of LTQ-Orbitrap profiling analysis of complex samples, the percentage of chimeras may reach as high as 50 percent of total spectra. These preliminary estimations of the frequency of mixture spectra justify the necessity of developing new method for characterizing those spectra. In addition, some new experimental modes also rendered the increasing necessity for peptide identification from mixture spectra. For instance in the data-independent acquisition (DIA) strategy [47], precursors over a large mass range are co-fragmented to avoid the sequencing issues existent in data-dependent acquisition strategy,

therefore the generation of mixture spectra is unavoidably. Other new acquisition strategies that tend to generate mixture spectra include the Multi-stage MS technique [48] and the SWATH acquisition strategy [49]. Peptides with similar masses and chromatographic properties can occur quite frequently and generates mixture spectra which will significantly reduce the identification efficiency due to existence of unidentifiable fragment ions derived from the co-eluting precursors. Most of the mainstream approaches now make the assumption that each MS/MS spectra comes from only one peptide, therefore they are not suitable for the identification of mixture spectra. New computational methods are necessary to handle the problem of mixture spectra identification. Such research will be very beneficial to the general objective of building proteome profile of a specific species, for instance, the complete human proteome profile.

The identification of mixture spectra is a challenging research topic. The ability to identify mixture spectra will be useful to alleviate the algorithmic bottleneck that exists in the current computational analysis of mass spectrometry based proteomics. To the best of our knowledge, the research on computational analysis of mixture MS/MS spectra remains unexplored. It is necessary and promising research to develop new algorithms and software tools for this objective. The work in this thesis focused on developing novel algorithms for peptide identification from mixture MS/MS spectra.

1.2 Chapter Outlines

The thesis is organized in the following chapters:

In chapter 1, we present a brief introduction to the background and motivation of the research.

In chapter 2, we provide the necessary fundamentals for mass spectrometry based pro-

teomics research, which include the biochemical basic knowledge, and the mass spectrometry technology. And in this chapter we also give a general summary on the current computational methods for mass spectra identification.

In chapter 3, we formulate the mixture spectra *de novo* sequencing problem mathematically, and propose a dynamic programming algorithm for the problem. Additionally, we use both simulated and real mixture spectra datasets to verify the merits of the proposed algorithm. Some materials of this chapter have been published in referred research articles in [50, 51].

In chapter 4, we propose an approach for matching mixture tandem mass spectra with a pair of peptide sequences acquired from the protein sequence database by incorporating a special *de novo* assisted filtration. The preliminary experimental results demonstrate the efficiency of the integrated filtration strategy in reducing examination space and verified the effectiveness of the proposed matching method. Some materials of this chapter have been published in peer reviewed paper in [52] .

In chapter 5, we conclude the whole thesis by summarizing the major content in the research and giving a discussion about the possible future research work.

Chapter 2

Fundamentals

2.1 Biochemistry Basics

Proteins are large biological molecules, consisting of one or more long chains of amino acids residue. Proteins are essential to a variety of biological functions within living organisms. Protein differs from one another primarily in their sequence of amino acids, which is translated from the DNA sequences of the species. The amino acids sequence can fold into some specific three-dimensional structure that will ultimately determine the biological activity of the protein. After translation, there exists another biological process called post-translational modification (PTM) which alters the protein sequence by attaching an amino acid residue with some biochemical functional groups. The PTM further extends the range of functions of the proteins in organisms.

2.1.1 Amino Acid and Protein

Amino acids are molecules containing both amine ($-NH_2$) and carboxylic acid ($-COOH$) functional groups, along with a side-chain specific to each amino acid. These molecules are of

particular importance in biochemistry, where the term *amino acid* is used to refer specifically to the alpha-amino acids with a general formula $H_2NCHRCOOH$ where R is a side chain. The general structure of an alpha-amino acid is shown in Figure 2.1.

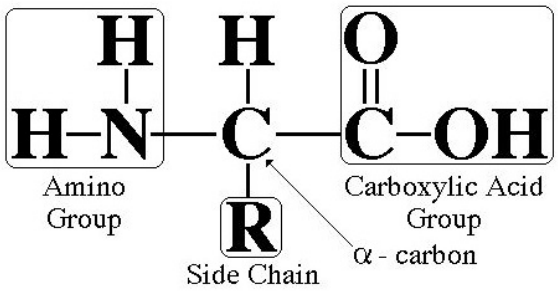

Figure 2.1: Structure of alpha-amino acid.

An amino acid removing one hydrogen atom (H) from the amine group (N-terminal) and one hydroxyl group (OH) from the carboxylic group (C-terminal) is referred as the *amino acid residue*. The general structure of an amino acid residue is shown in Figure 2.2.

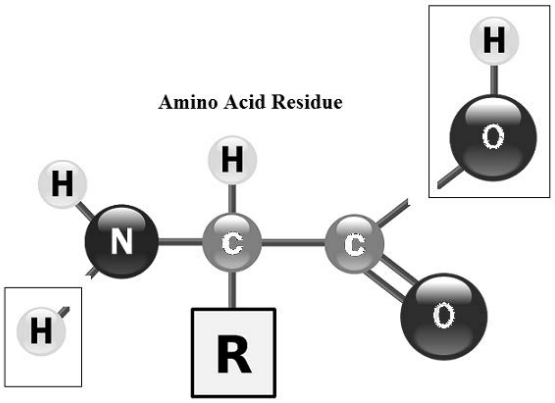

Figure 2.2: General structure of an amino acid residue [53].

There are 20 common alpha-amino acids. The difference of them only lies in the structure of the side chain groups (R groups). In Table 2.1, it contains the mass and composition information regarding all the 20 common amino acid residues.

Table 2.1: Residue Mass and composition of the 20 standard amino acids.

Name	3-letter Symbol	1-letter Symbol	Monoisotopic Mass	Average Mass	Residue Composition
Alanine	Ala	A	71.03711	71.08	C_3H_5NO
Arginine	Arg	R	156.10111	156.2	$C_6H_12N_4O$
Asparagine	Asn	N	114.04293	114.1	$C_4H_6N_2O_2$
Aspartic Acid	Asp	D	114.02694	115.1	$C_4H_5NO_3$
Cysteine	Cys	C	103.00919	103.1	C_3H_5NOS
Glutamic Acid	Glu	E	129.04259	129.1	$C_5H_7NO_3$
Glutamine	Gln	Q	128.05858	128.1	$C_5H_8N_2O_2$
Glycine	Gly	G	57.02146	57.05	C_2H_3NO
Histidine	His	H	137.05891	137.1	$C_6H_7N_3O$
Isoleucine	Ile	I	113.08406	113.2	$C_6H_{11}NO$
Leucine	Leu	L	113.08406	113.2	$C_6H_{11}NO$
Lysine	Lys	K	128.09496	128.2	$C_6H_{12}N_2O$
Methionine	Met	M	131.04049	131.2	C_5H_9NOS
Phenyalanine	Phe	F	147.06841	147.2	C_9H_9NO
Proline	Pro	P	97.05276	97.12	C_5H_7NO
Serline	Ser	S	87.03203	87.08	$C_3H_5NO_2$
Threonine	Thr	T	101.04768	101.1	$C_4H_7NO_2$
Tryptophan	Trp	W	186.07931	186.2	$C_{11}H_{10}N_2O$
Tyrosine	Tyr	Y	163.06333	163.2	$C_9H_9NO_2$
Valine	Val	V	99.06841	99.13	C_5H_9NO

And the structures of 20 standard alpha-amino acids are listed in Figure 2.3 [54].

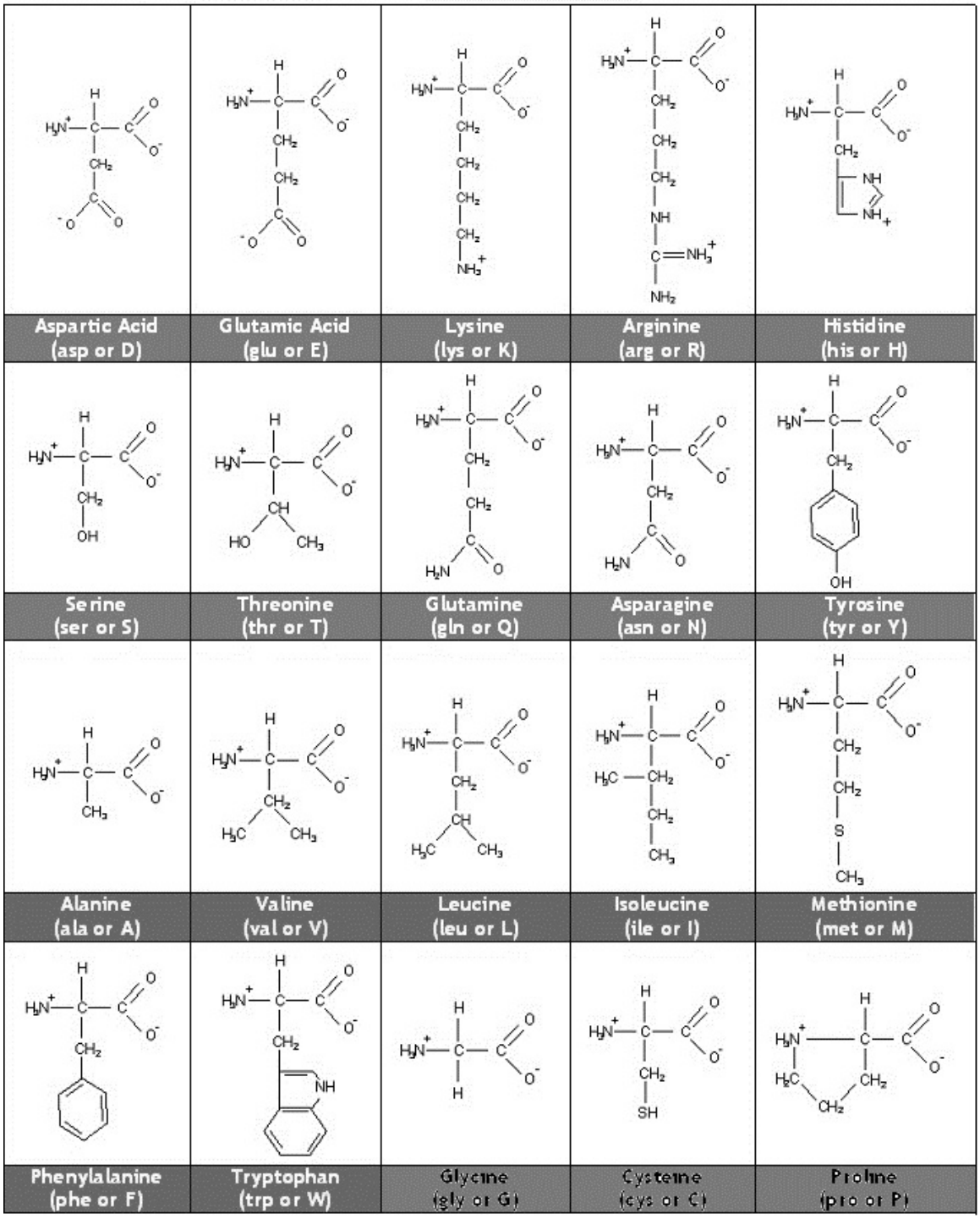

Figure 2.3: Chemical Structure of 20 standard alpha-amino acids.

Amino acids are the structural units that constitute polypeptides and proteins. They join together to form short polymer chains called peptides or longer chains called proteins. The

reaction process that two amino acids to form a peptide bond is called condensation [55]. The peptide bond (amide bond) is a covalent chemical bond formed between two amino acid molecules. In condensation, two amino acids approach each other, with the acid moiety of one coming near the amino moiety of the other. One loses a hydrogen and oxygen from its carboxyl group (COOH) and the other loses a hydrogen from its amino group (NH_2). This reaction produces a molecule of water (H_2O) and two amino acids joined by a peptide bond. The condensation of amino acids to form a peptide bond is shown in Figure 2.4.

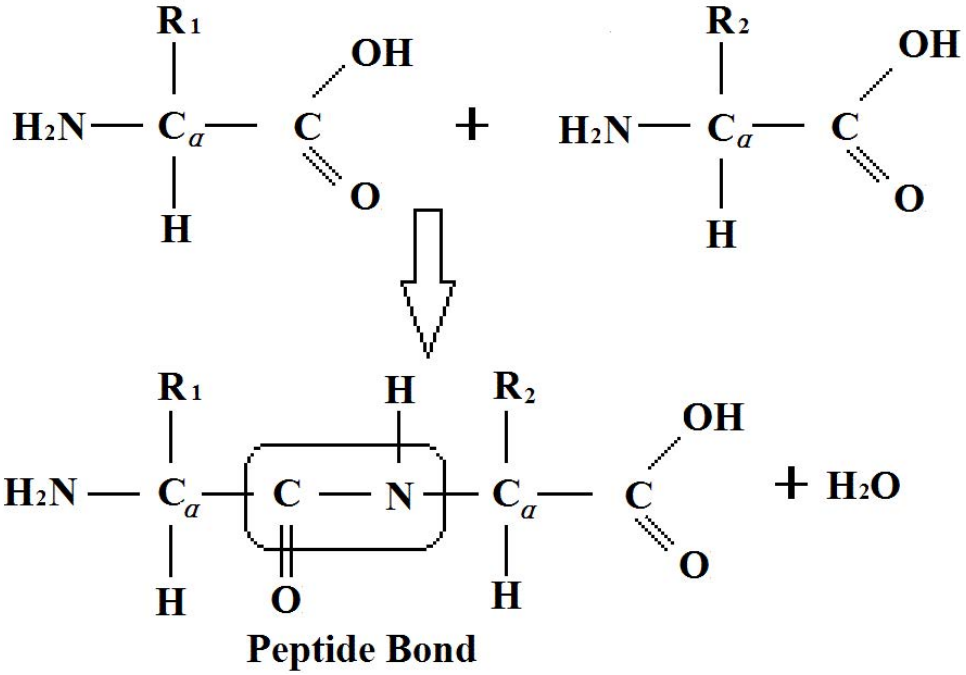

Figure 2.4: Condensation of two amino acids to form a peptide bond

The primary structure of a protein is a linear chain of amino acids, and generally proteins contain at least 40 amino acid residues [56]. Figure 2.5 shows the protein primary structure contains sequence of a chain of amino acids. Once the chain of amino acids has been assembled by the ribosome, it tends to fold in on itself. The function of a protein is directly dependent on its 3-dimensional structure. Remarkably, the higher level of protein structure is determined by the sequence of amino acids in the protein polymer [57]. The weak hydrogen bonds between peptide bonds in different parts of the polypeptide will bring the secondary structure to the

protein. Several types of amino acids contain sulphur in their side chain groups, for instance the amino acid *Cysteine* containing a sulphur atom in its R-group, tend to form disulphide bonds between them. This will result in the complex 3-dimensional tertiary structure of the protein. More than one polypeptide chains folding and connecting together will eventually bring the quaternary structure for some proteins.

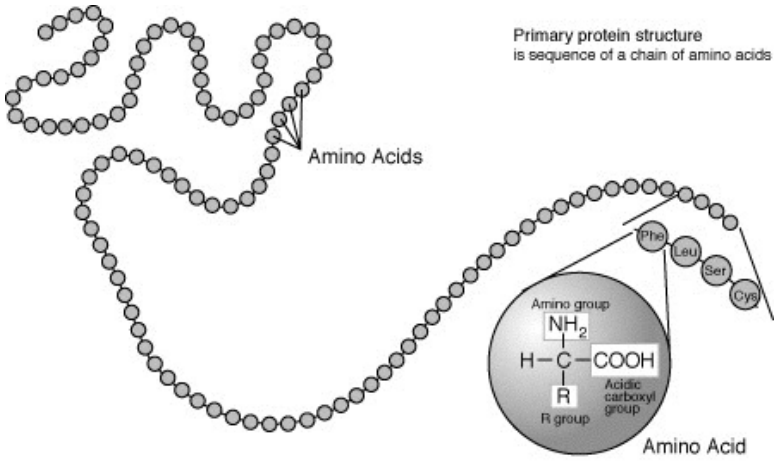

Figure 2.5: The primary structure of protein

2.1.2 Post-Translational Modifications

Post-translational modifications (PTM) of protein increase the functional diversity of the proteome by the covalent addition of some molecular groups. PTMs are chemical modifications that play a key role in the biological functions of the altered proteins, because they adjust the physical and chemical properties, stability and activity of a protein, thus altering the structure and function of a protein [58, 59, 60]. Most eukaryotic proteins are post-translationally modified proteins [61], and PTMs have also been frequently reported to be involved in various diseases including cancers and heart disease [62]. Today, there are hundreds of post-translational modifications reported by previous study, in which, the Unimod PTM database contains more than 500 entries [63] and the DletaMass database includes over 300 entries [64]. The most frequently observed PTMs includes phosphorylation, glycosylation, methylation, acetylation and

acylation [65]. Table 2.2 contains the summary of 25 Post-Translational Modifications which are frequently reported by previous experiments [66].

Table 2.2: Mass and modification residue of some commonly reported PTMs

Entry No.	Mass(Δm)	Residue	Modification Name
1	-48.003372	M@C-term	Homoserine lactone
2	-29.992805	M@C-term	Homoserine
3	-18.010565	C@N-term	Dehydration
4	-18.010565	E@N-term	Pyroglutamic Acid from E
5	-17.026548	Q@N-term	Pyroglutamic Acid from Q
6	-0.984016	[X]@C-term	Amidation
7	0.984016	R,N,Q	Deamidation
8	14.01565	E,[X]@C-term	Methylation
9	15.994915	W,H,C,M	Oxidation or Hydroxylation
10	21.981943	D,[X]@C-term	Sodium adduct
11	27.994915	T,[X]@N-term	Formylation
12	31.989828	M	Dihydroxy(Di-oxidation)
13	42.010567	C,[X]@N-term	Acetylation
14	43.005814	K,[X]@N-term	Carbamylation
15	44.026215	C	Ethanolation
16	45.98772	C	Beta-methylthiolation
17	57.021465	C	Carbamidomethyl
18	58.005465	C	Iodoacetic acid derivative
19	71.03712	C	Acrylamide adduct
20	79.95682	Y,T,S	O-Sulfonation
21	79, 96633	D,Y,H,T,S	Phosphorylation
22	99.06841	C	N-sopropylcarboxamidomethyl
23	105.057846	C	S-pyridylethylation
24	162.0528	Y,[X]@N-term	Hexoses
25	203.0794	S,T,N	N-Acetylhexosamine
26	210.1937	K,[X]@N-term	Myristoylation
27	226.07759	K,[X]@N-term	Biotinylation

2.2 Mass Spectrometry Technology

Mass spectrometry (MS) is an analytical technique for discovering the composition of a sample by measuring the mass values of the molecules in it. It has been used for different analytical purposes, both qualitative and quantitative analysis, including identifying the composition and structure of the target compounds and measuring the abundances of interested molecules. Nowadays, mass spectrometry has been widely used to analyze the physical, chemical or biological properties of a large amount of compounds [67].

2.2.1 Instuments and Configuration

A mass spectrometer typically contains three components: the ionization source, the mass analyzer and the detector. Figure 2.6 shows the schematic of the basic components of a mass spectrometer. Molecules are first ionized in the ionizer, then the ions are separated according to their different m/z in the mass analyzer, and finally the separated ions are detected in the detector to form an MS spectrum which is comprised of a series of peaks.

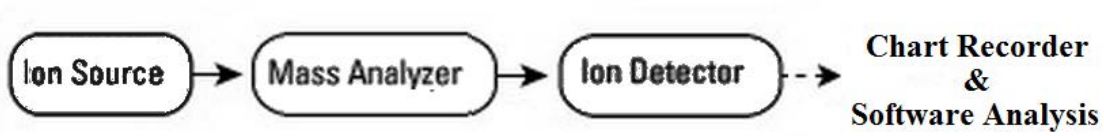

Figure 2.6: Schematic of the basic components of a mass spectrometer including *Ion Source*, *Mass Analyzer* and *Ion Detector*.

Figure 2.7 shows the schematic of a time-of-flight (TOF) mass spectrometer. Charged ions from the ionization source are accelerated into the TOF tube, which contains an electric field free flight region. The kinetic energy gained during acceleration decreases with increasing mass, such that heavier ions will fly slower and have a longer time-of-flight. This is the basis of TOF mass analysis.

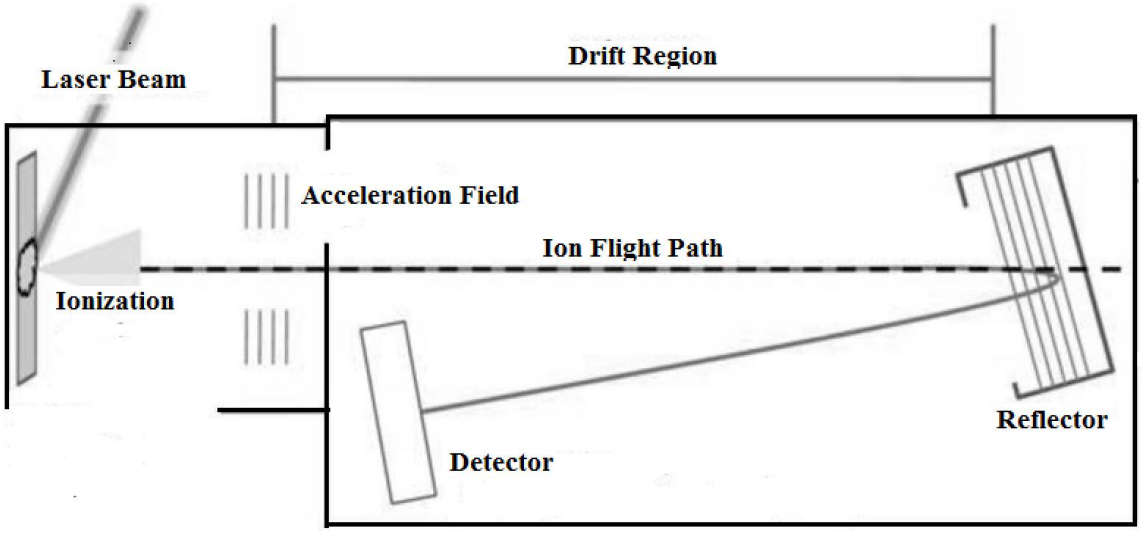

Figure 2.7: Basic principles of time-of-flight (TOF) mass spectrometry analysis. Following ionization, sample ions are accelerated into an electric field free region for drift. The larger the ion, the less energy it will gain during acceleration and as a result it will travel slower than smaller ions [68].

2.2.1.1 Ionization Source

In the ionizer, compounds in the sample are ionized by some ionization technique, for instance by impacting the compounds with electron beam, which renders the formation of charged molecules. Two types of ionizers are commonly used in mass spectrometry based proteomics study, the MALDI (Matrix-Assisted Laser Desorption/Ionization) and ESI (Electrospray Ionization). In MALDI, the analytes of interest is uniformly mixed with a large quantity of matrix material which is comprised of certain kind of low molecular weight Ultraviolet absorbing substance, then a pulsed laser irradiates the matrix-sample spot, triggering desorption and vaporization of the sample and the matrix. The matrix absorbs the UV laser energy, preventing the sample from being destroyed, also causes it to dissociate and will typically transfer a proton to the molecule to facilitate the ionization of the sample before they are accelerated into mass spectrometers [69]. Figure 2.8 shows how the ionization happens in MALDI.

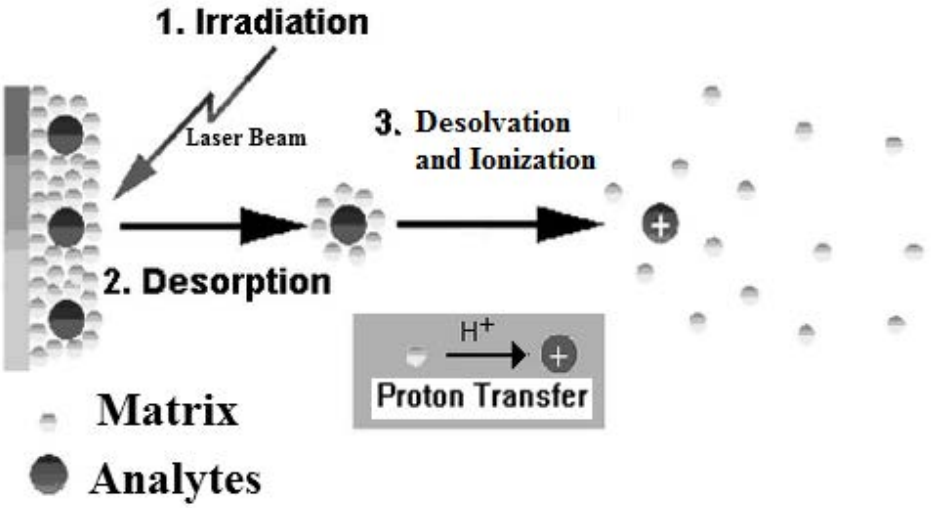

Figure 2.8: Matrix-Assisted Laser Desorption/Ionization (MALDI) [70]

In ESI, the liquid containing the analytes of interest is first dispersed into a fine aerosol before going through a capillary tube carrying high voltage, then the droplet is heated to aid the solvent evaporation which will subsequently makes itself smaller and denser charged. When reaching the Rayleigh limit, the surface tension that holds the droplet together will be surpassed by the electrostatic repulsion of like charges, this will cause the droplet to break into smaller yet stable droplets. The new droplets undergo similar desolvation and fission process which will eventually turn into a stream of charged ions [71]. Figure 2.9 shows the principle of Electrospray Ionization (ESI).

The primary difference is that MALDI produces singly charged ions ($z = 1$) and ESI produces singly and multiply charged molecules. One advantage of ESI is that a relatively large molecule can still be measured and profiled in the mass spectrometers when the charge state $z > 1$. Another advantage of ESI over other ionization method is that the instrument can be calibrated in the low m/z range, because of the multi-charge phenomena of the same ions. After being ionized, the molecules of interest will be transmitted into the electromagnetic filed of a mass spectrometer in which particles with different m/z values will demonstrate different motions when passing through the electromagnetic filed.

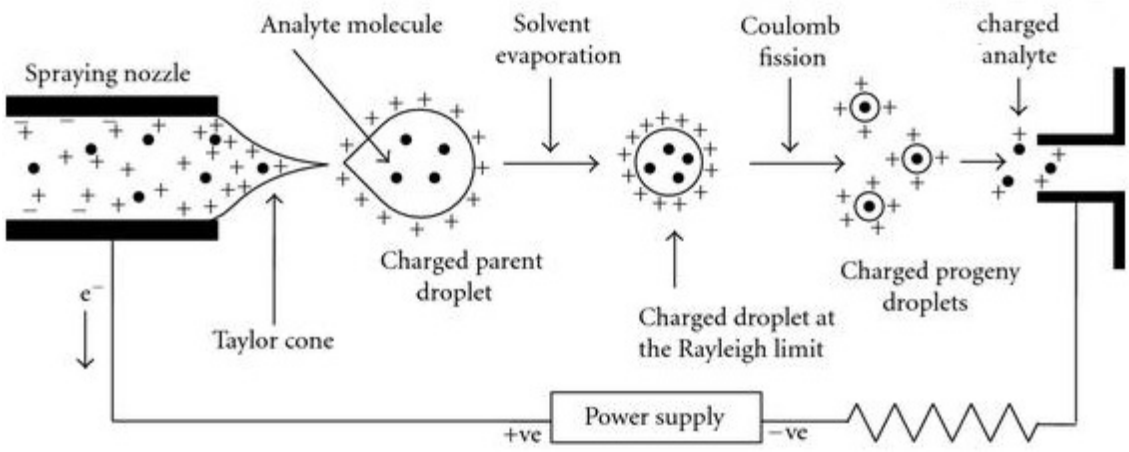

Figure 2.9: General Framework of Electrospray Ionization (ESI) [72]

2.2.1.2 Mass Analyzer

The mass analyzer is the instrument that separates the ions according to their mass-to-charge ratios(m/z). For mass analyzer, we consider three primary parameters: mass range, mass resolution and mass accuracy. The mass range of a mass analyzer determines the limits of m/z over which it can measure ions passing through. Only ions that fall into this range can be profiled. The m/z range used for proteomics analysis is typically from around 100 *Da* to a few thousand *Da* in which one *Da* is $\frac{1}{12}$ of the mass of a carbon atom (^{12}C), and is approximately the mass of one hydrogen atom. This mass range enables a balance between sensitivity and accuracy. In some mass spectrometry instruments, the mass range can be configured to span a rather large m/z range by trading off its resolution and mass accuracy. Also for the ion trap instruments, ions with low m/z values usually will not form peaks in the MS spectra due to the significant low m/z cut-off.

Mass resolution evaluates the ability of the mass spectrometer to distinguish two peaks with a slightly different m/z values. It is conventionally defined as $R = \frac{M}{\Delta M}$ (also called the

Resolving Power) in which the mass difference ΔM can be defined in different ways. In the valley definition, ΔM is the closest spacing between two peaks of equal intensity with the valley between them less than a specified fraction of the peak height. Typical fraction are 5%, 10% or 50%. While in the peak width definition, the value of ΔM is the width of the peak measured at a 50% of the peak height, which is also called the Full Width at Half Maximum (FWHM). Figure 2.10 shows the two definitions of mass resolution.

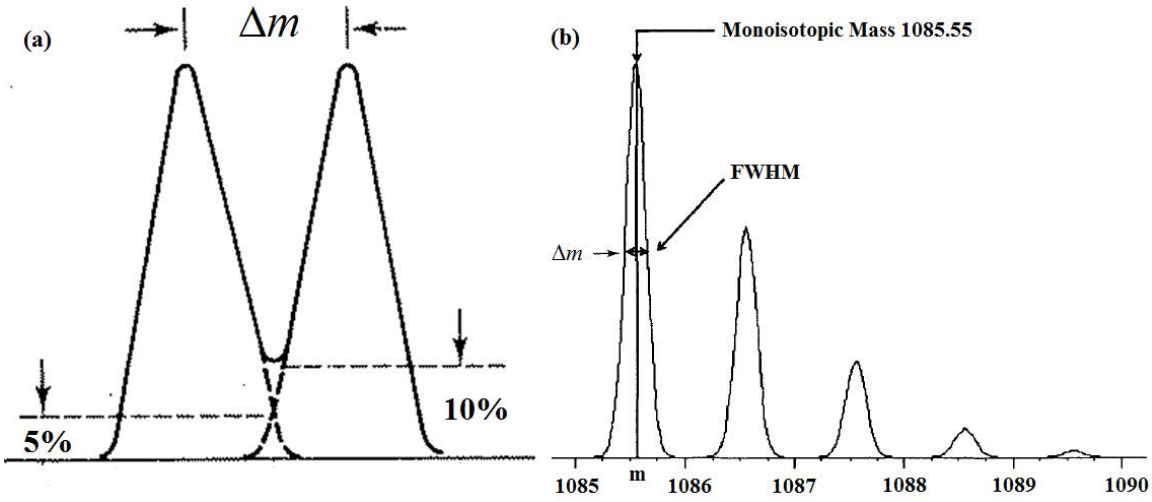

Figure 2.10: Examples for mass resolution based on two definition: (a) The valley definition of measuring peak separation. (b) The peak width definition of ΔM under the Full Width at Half Maximum (FWHM) scheme.

Larger resolution always indicates a better separation of peaks profiled in a mass spectrum. High resolution is considered to be with *Resolving Power* $RP \geq 5000$, which will greatly facilitate high precision measurements. Figure 2.11 shows an example of how the resolving power can have a dramatic effect on resolving isotopes. The mass spectrum of a protonated molecule is obtained at different resolving powers of 200, 2000 and 20000 under the FWHM definition of resolution, from which we can see that more details can be observed under a higher resolution mode.

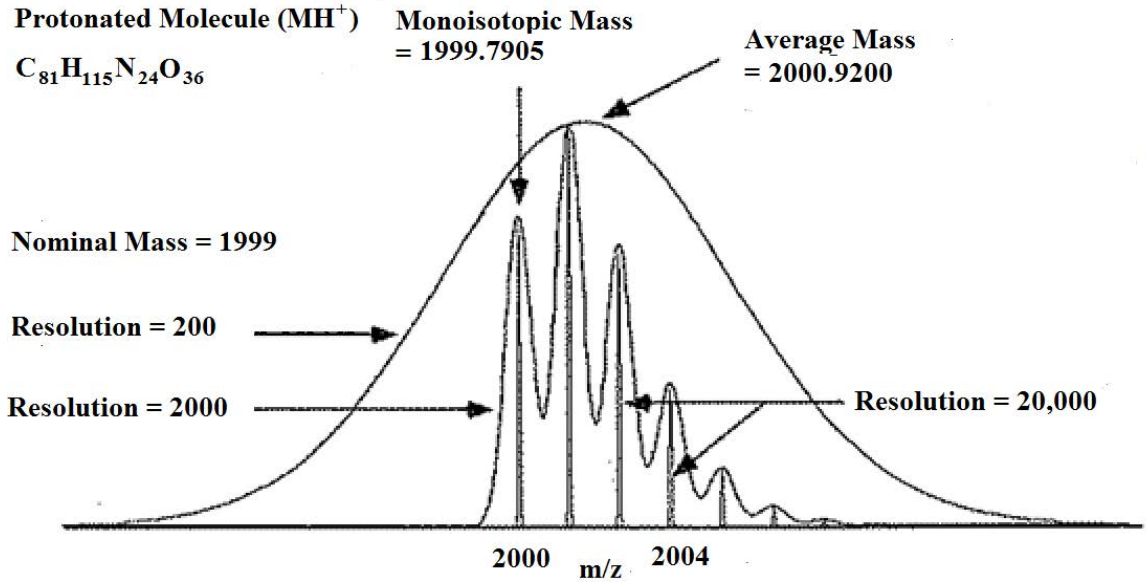

Figure 2.11: Achievable measurement under different resolving power[81]. Drastic difference exists in the shape of the same signal after measuring under different *Resolving Power*.

Mass accuracy defines how much accuracy of the mass value a mass analyzer can provide. It is normally measured by *millimass unit* (*mmu*) or *parts per million* (ppm). A *mmu* is equivalent to $1/1000$ of the *unified atomic mass unit* (u). Now the *unified atomic mass unit* (u) is displaced by the unit *dalton*, so 1 *mmu* equals to 1 *millidalton* (*mDa*). Mass accuracy expressed in ppm is calculated in the following way: $A_{ppm} = \frac{(|m_1 - m_2|)}{m_2} \times 10^6$ where m_1 is the real mass, and m_2 is the mass given by the mass spectrometer. So given the mass accuracy A_{ppm} and the mass value of an ion MW, we can covert the accuracy back to the unit of *mDa* by the formula $A_{mDa} = \frac{A_{ppm} \times MW}{10^3}$.

Several types of mass analyzers have been developed, including the Quadrupole mass analyzers [73], Ion trap analyzers (Quadrupole ion trap, QIT [74]; Linear ion trap, LIT/LTQ [75]), Time-of-flight (TOF) analyzers [76], Fourier transform ion cyclotron resonance (FT-ICR) [77], and Orbitrap [78]. Each type of mass analyzer has different capabilities in terms of sensitivity, accuracy, resolution, m/z range and some other properties [79, 80]. Table 2.3 provides a summary of the performance characteristics for each mass analyzer.

Table 2.3: Comparison of the typical performance characteristics of several types of commonly used mass analyzers including *Resolution*(represented by *Resolving Power*), *Accuracy*, m/z *range* and *Scan Rate*.

Mass Analyzer	Resolving Power	Accuracy(ppm)	m/z Range	Scan Rate
Quadrupole	1,000	100-1,000	50-2,000; 200-4,000	Moderate
QIT	1,000	100-1,000	10-4,000	Moderate
LTQ	2,000	100-500	50-2,000; 200-4,000	Fast
TOF	10,000-20,000	10-100	No upper limit	Fast
FT-ICR	100,000-750,000	<2	50-2,000; 200-4,000	Slow
Orbitrap	30,000-100,000	2-5	50-2,000; 200-4,000	Fast

2.2.1.3 Ion Detector

After the ions are separated by the mass analyzer, they reach to the ion detector. The detector records the current signal produced by an ion impinging event or the charge induced when an ion passing by. Generally when an ion hits the metal surface of the detector, its charge will be neutralised by an electron emitted from the surface onto the ion. The kinetic movement of electrons forms an electric current, which will be recorded subsequently. Because the number of ions leaving the mass analyzer at a specific instant is normally small, therefore it is significant to apply amplification technique following the detection to get a signal.

The simplest ion detector is the *Faraday Cup* which consists of a metal cup that collects all ions leaving the mass filter. The cup-shape metal surface enables itself to capture the secondary electrons emitted upon an ion impact event. The current flowing away from the Faraday cup will be recorded as a signal for further analysis. The Faraday cup is relatively low in sensitivity and slow in response time. Perhaps the most commonly used detector is called the *Electron Multiplier* which transfers the kinetic energy of incident ions to a dynode surface that in turn generates secondary electrons. A electron multiplier is normally comprised of a series of dynodes maintained at increasings potentials. An emission of electrons is caused by an ion striking the first dynode surface, and these electrons are then attracted to the next dynode with a higher potential and therefore more secondary electrons are emitted. Ultimately, as more dynodes are

involved, a cascade of electrons is collected and converted into a voltage signal that is proportional to the number of impinging ions. Another type of electron multiplier contains just one continuous dynode instead of several discrete dynodes. The principle is similar where the electrons are multiplied during several consecutive impacts inside one dynode. The advantage of an electron multiplier in contrast to a Faraday cup is the high amplification factor and the fast response time. Figure 2.12 shows the schematics of the mentioned ion detectors.

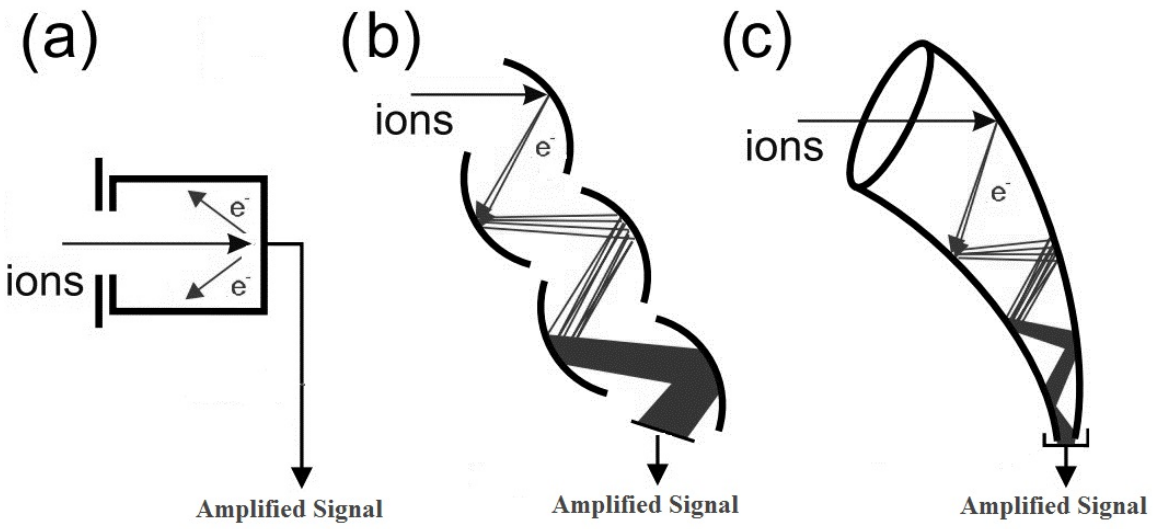

Figure 2.12: Ion detectors: (a) Faraday cup, (b) Electron multiplier with discrete dynodes maintained at particular potentials and (c) Electron Multiplier by a series of consecutive impinging events within one continuous dynode [82].

2.2.1.4 Mass Spectrum

Mass spectrometers are normally connected to computers with software that analyze the ion detector data and produce graphs that organize the detected ions by their individual m/z values and relative abundance. Ions with the same m/z will form a peak in the spectrum, and the intensity of that peak indicates the number of such ions observed by the detector which is directly related to the abundance of the corresponding ions in the sample. Because of the existence of element isotopes, there will also be isotopic peaks observed in the spectrum, from

which the charge state of the ion can be calculated. Figure 2.13 shows an example of a mass spectrum.

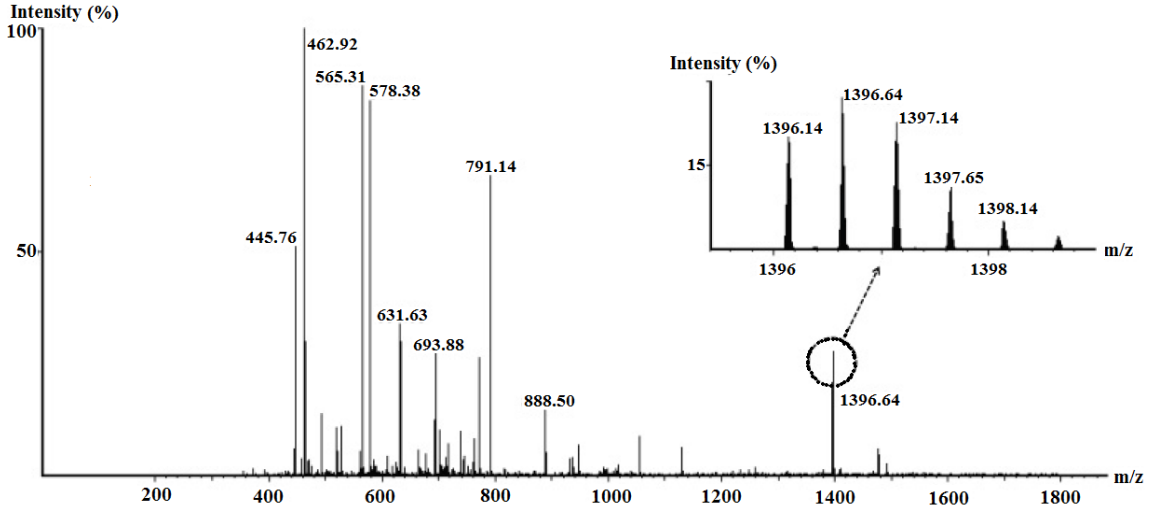

Figure 2.13: Example of a mass spectrum and the isotopic distribution [83]. After zooming in at a specific peak, the inner image shows more details about the adjacent isotopic distribution. Particularly, each peak actually spans a width on the m/z axis.

The inner image in Figure 2.13 shows the details of the isotopic distribution when zooming in a peak with $m/z = 1396.64$, in which we can find the corresponding monoisotopic peak with $m/z = 1396.14$. The charge state of the ion can be determined by the adjacent isotopic peaks [84, 85]. The adjusted monoisotopic m/z value can be further used to accurately interpret the mass spectrum [86, 87]. In particular, we can see that each peak indeed spans a bit width on the horizontal direction. Signal peaks will undergo a special process of *centroiding* before they are ready for computational analysis. During such process, each peak is assigned with a single m/z value, which usually represents the centroid of the peak shape.

2.2.2 Tandem Mass Spectrometry

Tandem mass spectrometry (MS/MS) has become the dominant method for proteomics study because it provides more information about a peptide than the traditional mass spectrometry

technique containing only a single round of mass analysis. It involves multiple stages of mass analysis which offers further information about specific ions. The first analyzer selects ions of interest at a specific m/z window which is normally a very small window only a few *Dalton*s wide. They are called the *precursor ions* (or *parent ions*). Then the ions are fragmented by some kind of dissociation method. And in the final step, the fragments are then separated based on their individual m/z ratios in another mass analyzer. Two types of mass spectra are generated in the tandem mass spectrometry experiment: *survey scan* (or *MS spectrum*) and *tandem mass spectrum* (or *MS/MS spectrum*). Figure 2.14 shows the general outline of tandem mass spectrometry.

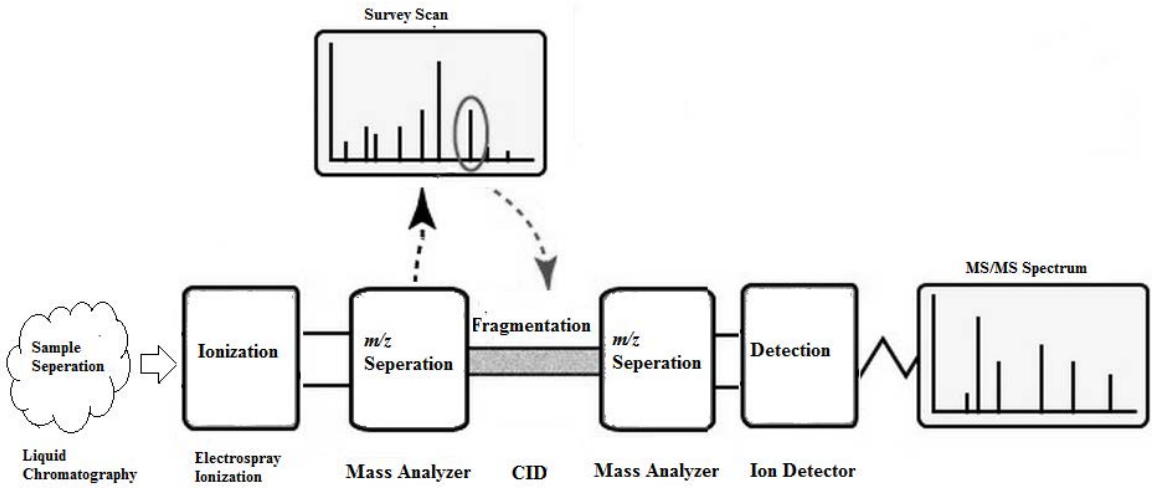

Figure 2.14: General diagram of tandem mass spectrometry. A sample is injected into the mass spectrometer, ionized and accelerated and then analyzed by the first mass analyzer. Ions of interests from the survey scan are then selectively fragmented and analyzed in another mass analyzer to form spectrum of the ion fragments. In the diagram above, it contains two separate mass analyzers, while in some other instruments it may contain two sequential analysis in the same mass analyzer.

2.2.2.1 Shotgun Proteomics

The applications of mass spectrometry in proteomics have drastically changed the characterization of proteins expressed in a cell or tissue from a labour-intensive style to a computational

assisted high-throughput fashion. Currently, *shotgun proteomics* refers to the utilization of *bottom-up* proteomics techniques in identifying proteins contained in complex mixtures using a combination of high performance liquid chromatography (HPLC) coupled with mass spectrometry technology [88, 89]. Figure 2.15 shows the typical procedure of characterizing peptides by current shotgun proteomics based approaches. The most common method of

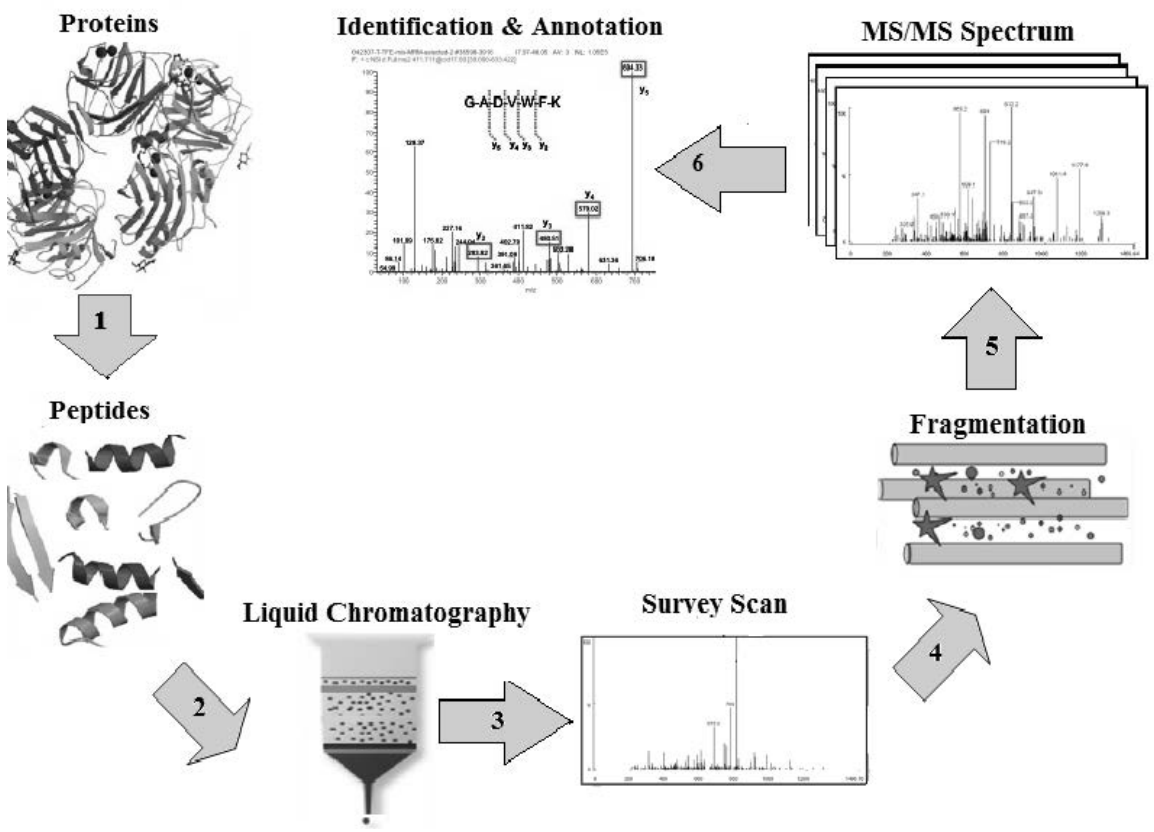

Figure 2.15: Workflow of shotgun proteomics by LC-MS/MS experiment. (1) Protein sample preparation by 2D-Gel and digestion by enzymes; (2) Peptide mixture separation by Liquid Chromatography; (3) Survey scan generation in tandem mass spectrometer; (4) Peptide fragmentation by HCD/ETD; (5) MS/MS spectrum generation in tandem mass spectrometer; (6) Computational analysis of mass spectrometry based proteomics data.

shotgun proteomics starts with the proteins in the sample being digested by single or multiple enzymes and the resulting peptides mixture are separated by liquid chromatography (LC), and then *tandem mass spectrometry* by nature serves as the prevalent method to identify and quantify peptides. Peptides are characterized from the collected mass spectral data by a variety of

different computational approaches, and proteins can be then inferred by matching these identified peptides to known protein sequences or assembling them into novel proteins [90, 91, 92]. Shotgun proteomics nowadays is widely used as the standard analytical approach in proteomics research.

2.2.2.2 Fragmentation and Ion Type

In a tandem mass spectrometry experiment, selected peptide precursors will be further fragmented before they are transferred to the second mass analyzer. Fragment ions observed in an MS/MS spectrum depend on many different factors including the primary sequence of peptide, the amount of internal energy, how the energy was introduced, charge state and so on. Theoretically, each fraction of the peptide will generate one peak in the resulting MS/MS spectrum. According to where the cleavage occurs, fragment ions can be classified into six basic types shown in Figure 2.16. Fragment ions retained with the *N-terminus* (the amino group) are $a-$, $b-$, and $c-$ion, while with the *C-terminus* (the carboxylic acid group) are $x-$, $y-$, and $z-$ion respectively. The subscript following each label indicates the number of amino acid residue kept in that fragment. If two backbone cleavages happen at the same time, internal fragment

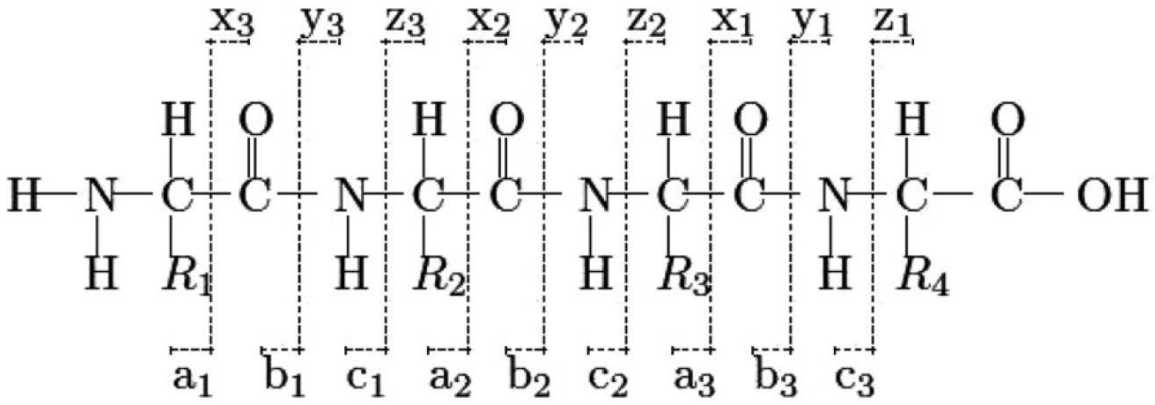

Figure 2.16: Six basic types of fragment ions, including $a-$, $b-$, $c-$, $x-$, $y-$, and $z-$ions, generated by cleavages on a peptide backbone structure with four amino acids.

ions will be generated. Two types of internal fragment ions are sometimes observed, including

the *amino-acylium* ions which are formed by a combination of *b* type and *y* type cleavages and the *amino-immonium* ions which are produced when a combination of *a* type and *y* type cleavages. *Immonium ion* is a special case of internal fragment with just a single side chain resulted from a combination of *a* type and *y* type cleavages. The immonium ions observed at the low end of some MS/MS spectrum which are usually considered as diagnostic ions that provide clues to the presence of specific amino acid residues in the peptide sequence. Though not often observed, the side chains of the precursor might also be fragmented which will generate several types of satellite ions, including the *d*−, *v*− and *w*−ions. Figure shows the structure of the internal fragment ions described above.

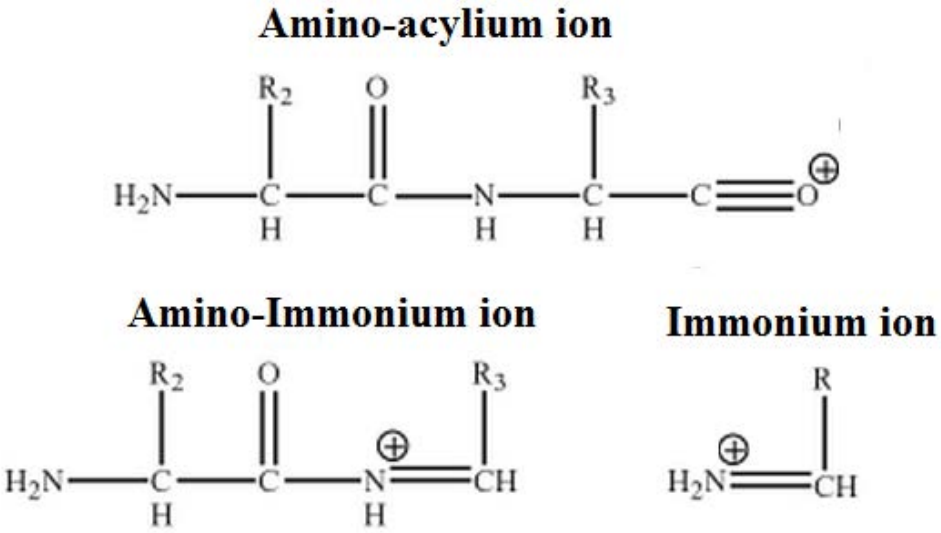

Figure 2.17: Structures of the internal fragment ions, including the *Amino-acylium* ion produced through a combination of *b*− and *y* type cleavages, the *Amino-immonium* ion formed by *a*− and *y*− type cleavages together, and the *Immonium* ion [93]

There are several fragmentation techniques available for tandem mass spectrometry, and each of those approaches tend to generate different types of ions. Three commonly used fragmentation methods in the current mass spectrometry based shotgun proteomics include the *Collision-Induced Dissociation (CID)* [94], the *Higher-energy Collisional Dissociation (HCD)* [95], and the *Electron-Transfer Dissociation (ETD)* [96]. The *Collision-Induced Dissociation* is also called *Collisionally Activated Dissociation (CAD)*, in which the precursor ions are first

accelerated by some electrical potentials to higher energy state and then collided with neutral molecule such as helium, nitrogen or argon. Part of the kinetic energy of the peptide ions is transferred to internal energy which induces the breaking of the peptide backbone. $b-$ion and $y-$ion are the most frequently observed ion types in the MS/MS spectra generated from CID fragmentation. The *Higher-energy Collisional Dissociation (HCD)* adopts a similar mechanism as CID fragmentation, from which the MS/MS spectra generated are also dominated by $b-$ions and $y-$ions. One advantageous aspect of HCD compared with CID is its ability to facilitate the generation of more $a-$ions and other smaller fragments which as a consequence will provide more information regarding the structure and composition of the molecule being analyzed. Another aspect worth mentioning is that HCD is specifically deployed with the *Orbitrap* mass spectrometers, therefore the MS/MS spectra obtained from HCD fragmentation always have more accurate m/z values. The *Electron-Transfer Dissociation (ETD)*, similar to the *Electron-Capture Dissociation* [97] induces cleavage of charged molecules, such as peptides or proteins, by transferring electrons to them. The most frequently observed ions from ETD/ECD fragmentation are $c-$ions and $z-$ions resulting from a cleavage at the $N - C_\alpha$ bond, and other ions derived from them, such as $(c-1)-$ions and $z'-$ions (sometimes referred as $(z+1)-$ions) are also observed frequently in the collected MS/MS spectral data [98].

2.3 Interpreting MS/MS Spectra

Mass spectrometry has been used for many applications, among which protein identification is by far the most mature application in proteomics study. Successful protein identification is considered to be the first step of the proteomics data analysis. In such application, the mass spectra containing the structural information are used to identify the sequenced peptides, in which the interpretation process always finds the best matching peptide for the target spectrum. Then in the following step, those peptides identified with high confidence are further

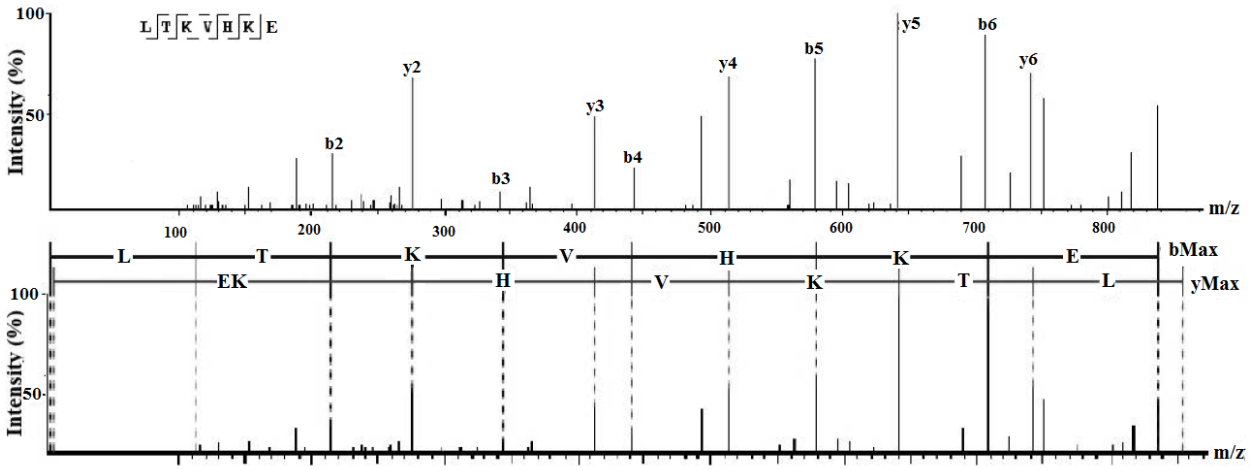

Figure 2.18: Example of an annotated spectrum of a successfully identified peptide **LTKVHKE**. In the upper spectrum, the theoretical $b-$ and $y-$ions match well with most of the significant peaks in the spectrum with small mass errors. In the nether spectrum, intervals of the explained peaks are labelled with the corresponding amino acids that have the same mass values.

assembled together to infer the proteins existing in the sample. Figure 2.18 shows an annotated spectrum of a successful identification. The process of reconstructing peptide sequence from mass spectral data is used to be done manually by biochemists. However, the amount of mass spectrometry data collected in the wet-lab experiment is growing drastically in recent years, which in turn makes it impractical for researchers to manually interpret the mass spectral data and the necessity of automated approaches to assign peptide sequences to spectra has become urgent. Till now, extensive research has been made to the development of new computational approaches for mass spectra interpretation. In general, the computational approaches for mass spectral data analysis fall into two categories: database search and peptide *de novo* sequencing.

2.3.1 Database Search

A database search approach requires the assistance of protein sequence database which is supposed to contain all the target proteins, and the computational task is to select the correct

proteins from the database. Generally, approaches for database search proceed in a very similar way. The experimental MS/MS spectra are compared with the theoretical spectra generated by applying certain cleavage rules to the peptide candidates acquired from a protein database. Peptide candidates are scored in a way that allows the peptide which best matches the spectral data to be reported. After that the proteins possibly contained in the sample are inferred from peptide matches. The general outline of a database search method is shown in Figure 2.19.

To find the best matched peptide sequence, normally a *Peptide-Spectrum Match (PSM)* score is calculated to describe the quality of a match between a candidate peptide and a given spectrum, and in this process, an effective scoring function is of primary importance for the peptide identification accuracy. Usually, the intensities and the number of matched peaks, as well as the mass errors of the matching are taken into account in the scoring function. Another factor to be considered in the scoring procedure is the fragment ion types because of the fact that certain type of mass spectrometer usually produces peaks with higher intensity for certain ion types.

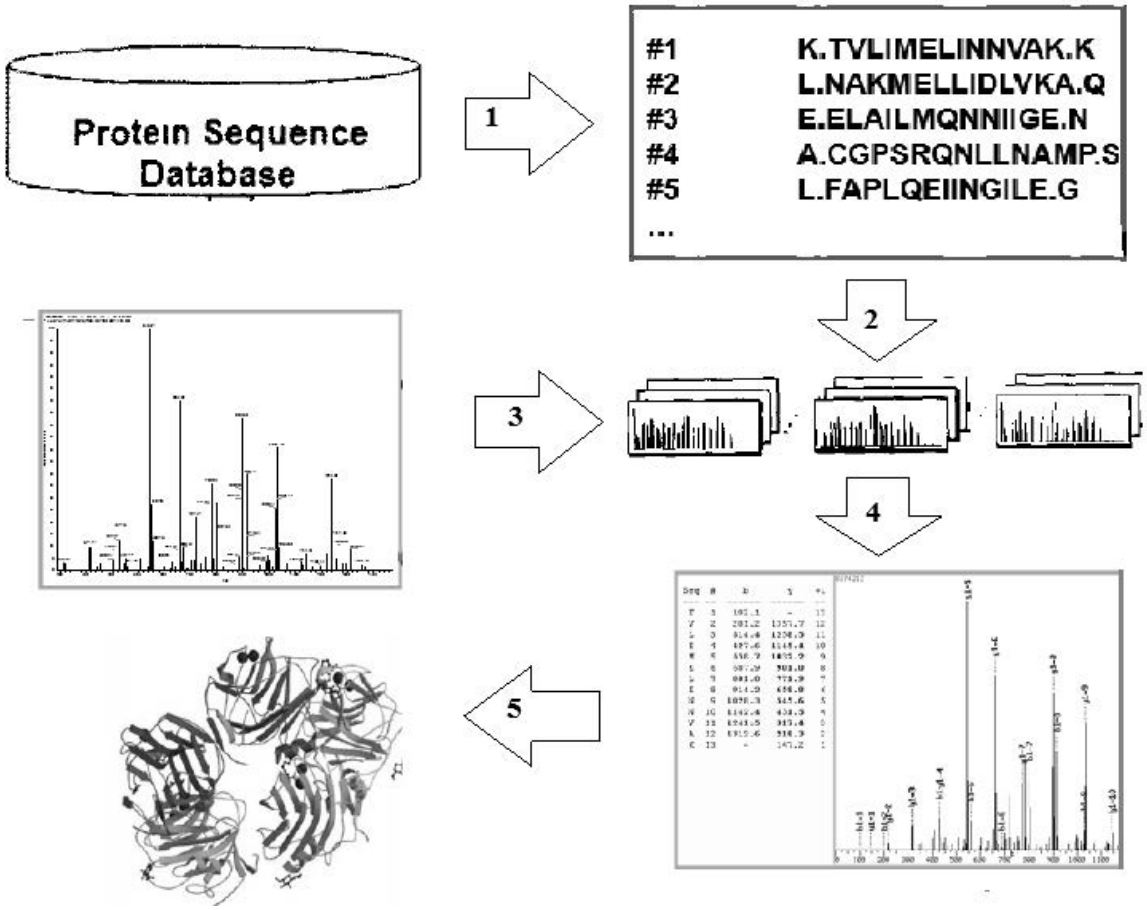

Figure 2.19: Workflow of database search approach. (1) Proteins are theoretically digested and peptide sequences that satisfy the precursor mass value of the input MS/MS spectrum are retrieved as the potential candidates. (2) Theoretical spectra are predicted from the selected peptides according to the certain fragmentation rules. (3) The experimental MS/MS spectra are compared with all the predicated spectra based on an appropriate scoring function. (4) Peptide candidates are ranked according to the scoring function, and the top-ranked candidates are outputted as the results. (5) The peptide matches are grouped together to identify the proteins in the sample.

After the identification results are reported, researchers always want to know the *False Discovery Rate (FDR)* at certain score threshold. Even equipped with proper scoring function, false discoveries still exist in the identification results due to different reasons [99, 100]. The FDR will provide researchers an extra dimension to evaluate which analysis results are trustworthy. Currently, the *Target-Decoy Database* method is widely used to validate the results by estimating the FDR. In such method, a *decoy* database is firstly constructed with similar statis-

tical properties as the target database, then the query spectra are searched against with both the target database and the decoy database. The false discovery rate at a specific score threshold is estimated by the number of matches in the decoy database with score above that threshold. Currently, there is still no unified method on how to construct the decoy database or even how to use the decoy database. A commonly used method is to reverse the protein sequences in the target database to get the decoy[101], while another research suggest that the decoy and the target database should be concatenated and searched together[102], and recently researchers have also proposed a two-round database search strategy coupled with a modified target-decoy database to estimate the FDR [10].

2.3.2 Peptide *De Novo* Sequencing

In proteomics, *de novo* sequencing is the process of deriving peptide sequences directly from tandem mass spectra without the assistance of a sequence database, therefore it is often used for novel protein identification or in the situation that the target protein or peptide is not included in the sequence database. It searches for the optimal combination of amino acids to match the input MS/MS spectra. Figure 2.20 shows the general workflow of peptide *de novo* sequencing. Such analysis has traditionally been performed manually by human experts, and more recently by computer programs that have been developed due to the requirement of high-throughput data analysis. Unlike the database search approach in which the algorithmic module is fairly straightforward, it can be as simple as enumerating every possible peptide sequences in the database with proper mass values, the *de novo* sequencing on the other hand requires more sophisticated algorithm component to seek the optimal solution. The naive solution of enumerating every possible amino acid combination for a given molecular mass value will take exponential time. One extensively studied mathematical model for *de novo* sequencing is to convert the spectrum to its corresponding *spectrum graph*, then the solution is to find an optimal path on the graph [15]. Figure 2.21 shows the general procedure of *de novo* sequencing by the

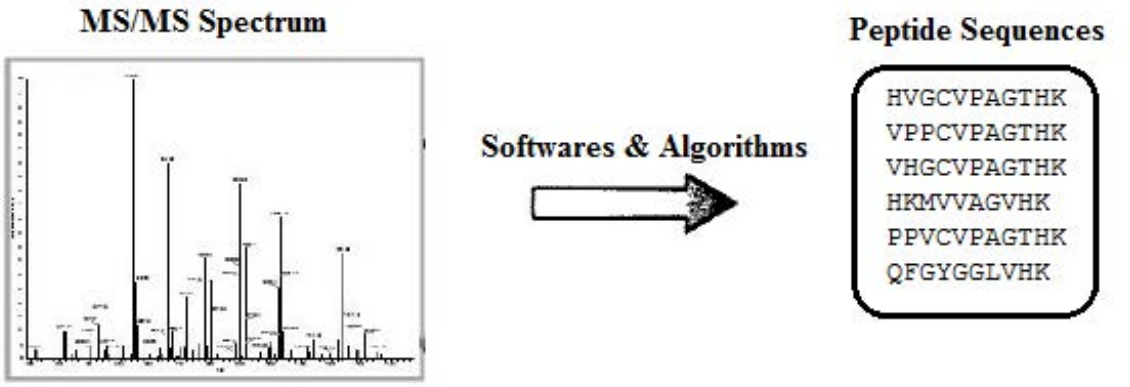

Figure 2.20: Workflow of peptide *de novo* sequencing. Usually, multiple sequences are constructed by the *de novo* algorithms, and always those sequences share some homologous parts.

spectrum graph model. Peptide *de novo* sequencing is considered to be a significantly harder problem than database search based identification, which requires much higher quality data in order to derive the complete sequence. A major disadvantage the *de novo* sequencing suffers is the low accuracy of the reported results, when the complete sequence is not able to obtain, it is supposed that partially correct sequence tags are also desired [16]. Those sequence tags can be used in assisting the database search method to reduce the examination space which will subsequently improve the sensitivity and accuracy of the identification. With the new advances in both algorithmic and instrumental aspects, the *de novo sequencing* software have made significant progresses in recent years, providing higher accuracy in identification and better coverage in protein sequence. In such scenario, novel protein sequencing becomes possible by combin-

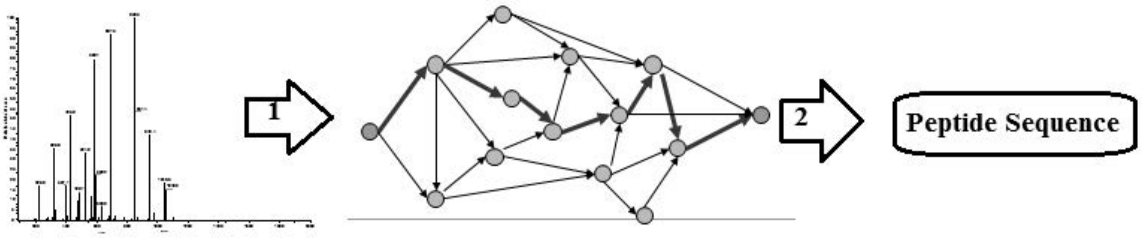

Figure 2.21: Schematic of the *graph spectrum* model. (1) The target MS/MS spectrum is converted to a spectrum graph in which each edge corresponds to an occurrence that the m/z difference of two peaks in the spectrum equals to the mass value of a certain amino acid residue. (2) Usually dynamic programming algorithms proceed in a carefully designated manner to find the path that represents the best solution.

ing the *de novo* sequencing method together with multiple enzymatic digestions of the proteins in the sample preparation phase [103, 104, 105, 106]. To alleviate the shortcomings of low quality spectra, another direction that researchers have been trying recently is to use multiple spectra of the same peptide generated by different fragmentation methods to improve the *de novo* sequencing accuracy [107, 108]. For instance, by combining the CID spectrum and ETD spectrum of the same peptide during the *de novo* procedure, the $b-$ and $c-$ion, as well the $y-$ and $z'-$ions can confirm each other and fill out the missing gaps that may occur in each individual fragmentation mode [109, 110, 111]. The combination of multiple fragmentation modes greatly increases the possibility of deriving a complete peptide sequence, as well reduces the misinterpretation of other peaks as ion ladder peaks.

2.3.3 Mixture Spectra Identification

The existence of mixture spectra significantly complicates the analysis of complex protein mixtures, yet their impact on the mass spectrometry based proteomics is still poorly understood. Now the proteomics research community is at the stage of realizing the significance of mixture spectra, and literature on interpreting mixture spectra in tandem mass spectrometry have already appeared. Masselon et al. described a method in [112] to identify the co-fragmented peptides by taking advantage of the extraordinary accuracy and resolution of the newly developed FT-ICR (Fourier Transform Ion Cyclotron Resonance) mass spectrometers. In this approach, groups of peptides from database are selected with high accuracy to generate the possible combinations from the co-sequenced precursors. Then the theoretical spectra of the co-fragmented peptides are compared with the acquired spectra to find the best matched combination. Zhang et al. introduced a database search engine, ProbIDtree [113], to identify co-eluting peptides from mixture tandem mass spectra. This method works in an iterative process of database searching in which ions assigned to a tentative peptide are subtracted from the acquired spectrum, and the remaining spectrum is used to detect another matched peptide. The software

then organize the tentative matched peptides in a tree structure and calculate an adjusted probability score to determine the correct identifications. Wang et al. proposed M-SPLIT [114], an MS/MS spectral library search software and demonstrated its potential to identify peptides from mixture spectra by matching the collected spectra with previously identified spectra. This method is limited in application due to the fact that it can only be used to identify peptide that has already been observed and confidently interpreted. Because of such dependency, the effectiveness of this method is highly subject to be impaired when experimental configuration changes. Recently, Wang et al. proposed another method, a database search tool, MixDB [115], which has the ability to interpret mixture spectra by using specifically designed scoring function for matching the mixture spectra with a pair of peptide sequences filtered from the protein database. The method explicitly formulates the occurrence of co-sequenced peptides in the same spectrum, and the FDR (False Discovery Rate) of the computational results is estimated as well. More recently, Zhang et al. presented a new workflow, DeMix [116], for the in-depth shotgun analysis of complex proteomics data. By capitalizing on the high resolution and mass accuracy of Orbitrap-based tandem mass spectrometry, the proposed method converts the unwanted co-fragmenting events into an advantage of interpreting chimeric tandem spectra generated by the co-fragmented peptides. In the meanwhile, there are several groups that conducted research on determining the accurate monoisotopic values of the co-sequenced precursors, and then applied the conventional database method to characterize the collected mass spectra in an iterative manner [117, 118, 119, 120].

Chapter 3

De Novo Sequencing of Mixture Spectra

3.1 Introduction

Due to many different reasons, such as incomplete genome sequencing, inferior gene prediction from the genome, or the sequence variations between two individuals of the same species, the target proteins may not be included in the sequence database, When this happens, *de novo* sequencing would become the only choice for identifying the peptides. For solving the problem of *mixture spectra de novo sequencing*, a reasonable assumption is that the computational methods should have the ability to derive multiple peptide sequences from a single mixture MS/MS spectra.

Besides all the shortcomings the conventional *de novo* sequencing methods suffer due to the imperfect data, there will be several new complications when coming up with the task of mixture spectra. First, we need to establish a valid mathematical model for mixture MS/MS spectra that are formed from co-fragmentation of multiple peptides. The precursors selected in the same isolation window in the MS survey scan may have significantly different intensities. The diversity of their intensities will greatly influence the peak intensities in the corresponding MS/MS spectra, therefore the model we set up should be able to adapt to the situation of

precursors co-sequenced with different abundances. Second, we have to design an effective scoring function to evaluate the similarity between the reconstructed peptides and the input mixture spectrum. For *de novo* sequencing, the scoring function must be designed as an inbuilt module in the efficient algorithm when computing the solution. The corresponding algorithmic module for mixture spectra *de novo* sequencing will have to take account of the fragments from both of the co-sequenced precursors simultaneously. Besides the dominant ions, such as $b-$ and $y-$ions present in the spectra, it is always required that other supporting ion types should also be considered when calculating the corresponding score. Third, we want to apply a reasonable method to deal with the overlapping peaks in the mixture MS/MS spectra. Theoretically, the overlapping peaks in MS/MS spectrum come from ions with similar m/z values. A possible tendency we can foresee is that the overlapping situation will be more complicated with mixture spectra, they may be induced by fragments from different peptides as well as fragments from the same peptides. Newly developed mass spectrometers can achieve very high resolution, for instance FT-ICR or LTQ-Obitrap. The accuracy of the measurement of the mass values could be as small as 1 ppm. Under such precision scheme, the occurrence of overlapping peaks is likely to be reduced, but still we can't neglect the influence of the overlapping peaks. Last but not least, we have to design efficient algorithm that enables to assign the significant peaks present in the mixture spectra to the appropriate candidate peptide sequences being constructed. Normally, *dynamic programming* is the choice for *de novo* sequencing algorithm. Particularly, the computation of one peptide sequence can be interfered by the presence of fragment peaks from another peptide. To alleviate the confusion of the computational engine, the algorithmic module should works in an carefully designated pathway to compute the solution.

Prior to this research, no algorithm has been reported for peptide *de novo* sequencing from mixture tandem mass spectra. Such study is necessary because the *de novo* sequencing can be regarded as a complementary method for spectra interpretation when the target peptide is not included in the database. Though computational methods for peptides *de novo* sequencing

always suffer the problem of low accuracy upon the derived peptide sequences, the partially correct peptide sequences can be still very useful for a subsequent database search to find the exact or homologous peptide. For instance, researchers have showed an increasing interest in combining *de novo* sequencing method with database search to more efficiently interpret MS/MS spectra lately [10, 121]. In those research, the *de novo* sequencing results that contain only partially correct sequences are searched against a protein database to further interpret the spectra. In the following section, we will formulate the problem of mixture spectra *de novo* sequencing mathematically and propose a dynamic programming algorithm to solve the problem. Then the effectiveness of the algorithm is evaluated by both simulated and real mixture spectra datasets.

3.2 Notations and Problem Definition

3.2.1 Basic Notations

Suppose that a mixture spectrum $\mathcal{M}$ is generated by the co-fragmentation of peptides P_1 and P_2, and $\mathcal{M}$ can be represented by a peak list $\mathcal{M} = \{(x_i, h_i)|i = 1, 2, ..., n\}$. Each element (x_i, h_i) represents a peak in the spectrum, in which x_i is the m/z value and h_i is the intensity of the peak. Data preprocessing including deconvolution is done by standard software packages which will convert the multiply charged peaks to their singly charged equivalents [122]. Therefore, in our research we assume that the mixture spectrum $\mathcal{M}$ only contains ions of charge one and the mass to charge ratio (m/z) of an ion is indeed equal to its mass value. Meanwhile, we use two molecular weight MW_1 and MW_2 to denote the precursor mass values of the two peptides which satisfy $|MW_1 - MW_2| \leq \Delta$. Δ is a small value predefined by the width of the mass spectrometer selection window. In the typical Data-dependent Acquisition (DDA) mode, the selection window of the mass spectrometer is usually a few *Dalton*(s) wide. The charge state of precursors is also considered as an influential factor. It is observed in experiment that

precursors with different charges are selected in the same isolation window and sequenced together. While in our research, we use a simplified model in the problem formulation that we assume the co-fragmented precursors take same charge state. In addition, we use Σ to denote the alphabet of 20 different types of amino acids. For an amino acid $a \in \Sigma$, we use $||a||$ to symbolize the mass of the amino acid residue. Here we have $\max_{a\in\Sigma} ||a|| = 186.08$ and $\min_{a\in\Sigma} ||a|| = 57.02$. It is worth mentioning that the two amino acids Isoleucine(I) and Leucine(L) have exactly identical mass values, so generally in *de novo* sequencing method, we consider them as the same amino acid.

Intuitively, a peak in $\mathcal{M}$ whose position matches with the mass of a fragment ion of peptide P is a positive evidence that $\mathcal{M}$ resulted from P. Similarly for a mixture spectrum, the more and higher peaks match with the theoretical fragment ions of P_1 and P_2, the larger likelihood that P_1 and P_2 are correct pair of peptides we are seeking. In this section we will formulate this intuition and model the mixture spectra *de novo* sequencing problem.

3.2.2 Mass Representation of Ion Fragments

In tandem mass spectrometry, peptides will be fragmented before they are transferred to the second mass analyzer. Theoretically, each fraction of the peptide will generate one peak in the MS/MS spectrum.

According to where the cleavage occurs, fragment ions can be classified into six basic types shown in Figure 2.16. Fragment ions retained with the N-terminus are $a-$, $b-$,and $c-$ion, while with the C-terminus are $x-$, $y-$,and $z-$ion respectively. The subscript following each label indicates the number of amino acid residue kept in that fragment. After fragmentation, peptide fractions are charged, and then followed by a measurement of the m/z and abundance values. Figure 3.1 illustrated the charging condition retained by each of the six basic ion types.

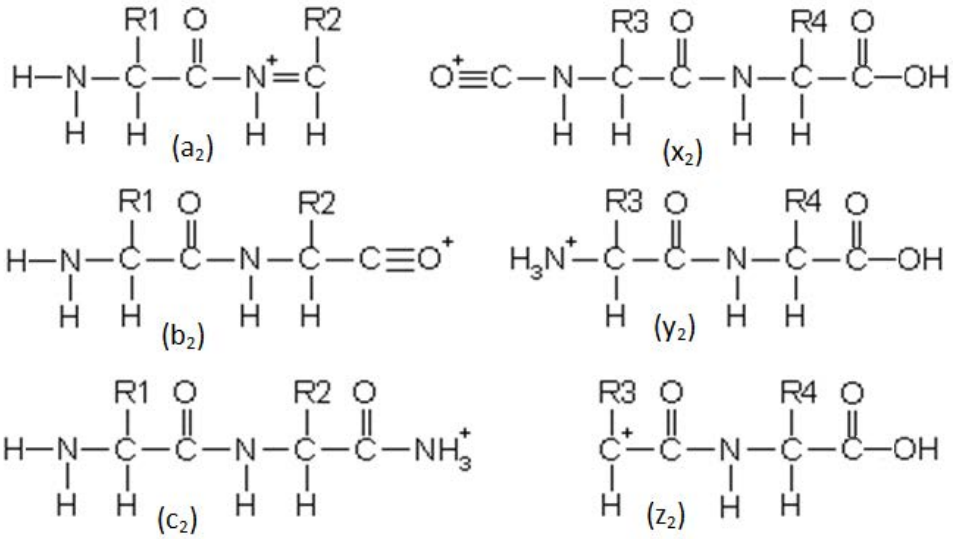

Figure 3.1: Charging condition of the six basic ion types after fragmentation. Particularly, in order to be properly charged, $y-$ion and $c-$ion will have to retain extra protons on some specific positions of the backbone structure.

Table 3.1 shows how to derive the mass of the fragment ions with respect to the molecular mass values of the neutral amino acid residues. The notation M_r represents the neutral mass value of the target ion. $[M]$ is molecular mass of the neutral amino acid residues. Table 3.2 contains the mass values and abundance level of several frequently used elements in calculating the mass of biomolecules. The mass of the element is represented by u or Da.[1]

[1]The *unified atomic mass unit (u)* or *Dalton (Da)* is the standard unit that is used for indicating mass on an atomic or molecular scale. One unified atomic mass unit is approximately the mass of one *nucleon* (either a single *proton* or *neutron*). It is defined as one twelfth of the mass of an unbound neutral atom of *Carbon-12* (^{12}C) and has a value of $1.660538921 \times 10^{-27} kg$.

Table 3.1: Neutral mass values of different ion types. To obtain m/z values, add or subtract protons as required to obtain the required charge and divide by the number of charges. For example, to get y^+, add 1 proton to the neutral mass value for y, then the actual mass value of the $y-ion$ with charge one is $OH + [M] + H + H$.

Ion Type	Neutral Mass (M_r)
a	$H + [M] - CHO$
a^*	$a - NH_3$
a^o	$a - H_2O$
b	$[M]$
b^*	$b - NH_3$
b^o	$b - H_2O$
c	$H + [M] + NH_2$
x	$OH + [M] + CO - H$
y	$OH + [M] + H$
y^*	$y - NH_3$
y^o	$y - H_2O$
z	$OH + [M] - NH_2$

Table 3.2: Mass values of isotopes. All the entries except the last one are the monoisotopic mass values of the corresponding elements.

Isotopes	Mass(u or Da)	Abundance(%)
^{12}C	12.000000	98.90
^{1}H	1.007825	99.986
^{14}N	14.003074	99.630
^{16}O	15.994915	99.762
^{32}S	31.972071	95.020
^{31}P	30.973762	100.000
^{13}C	13.003354	1.100

Let $P = a_1a_2...a_k$ be the string of amino acids, we define the residue mass of the peptide as $||P|| = \Sigma_{1\leq j\leq k}||a_i||$ and the actual mass of the peptide as $||P|| + ||\mathrm{H_2O}||$. Denote b_i and y_i to be the mass of the $b-$ion and $y-$ion of P with i amino acids respectively. From Figure 3.1, we know that $b_i = 1 + \Sigma_{1\leq j\leq i}||a_j||$. Similarly for the corresponding $y-$ion with the remaining $k - i$ amino acid residue, the mass can be computed with $y_{k-i} = 19.02 + \Sigma_{i+1\leq j\leq k}||a_j||$. Therefore we have $b_i + y_{k-i} = ||P|| + 20.03$.

Let $\Pi = \{y, b, a, c, x, z, y^*, y^o, b^*, b^o\}$ be all the ion types that we consider throughout the

paper. Assume x is the mass value of a $b-$ion, from Figure 3.1, we know that the masses of the $a-$ion and $c-$ion at the same location are $x - 27.99$ and $x + 17.03$ respectively. Meanwhile a $b-$ion may lose an ammonia to generate new ion of mass $x - 17.03$, or lose a water to form another ion of mass $x - 18.01$.

we use $\mathcal{B}(x)$ to denote the set of all the ion masses corresponding to this $b-$ion, then we have:

$$\mathcal{B}(x) = \{x, x - 17.03, x - 18.01, x + 17.03, x - 27.99\}. \tag{3.1}$$

Similarly, for each $y-$ion with mass x, we will have the following notation,

$$\mathcal{Y}(x) = \{x, x - 18.01, x - 17.03, x + 25.98\}. \tag{3.2}$$

that represent all the ion masses related to this $y-$ion. Reason that $\mathcal{Y}(x)$ has one fewer element than $\mathcal{B}(x)$ is because $y-$ion losing an ammonia and its corresponding $z-$ion are both $||y|| - 17.03$.

Theoretically, the spectrum of the co-fragmented peptides $P_1 = a_1a_2...a_n$ and $P_2 = b_1b_2...b_m$ should contain a peak at each of the following mass values:

$$S(P_1, P_2) = \bigcup_{i=1}^{n-1}[\mathcal{B}(b_i^1) \cup \mathcal{Y}(y_i^1))] \bigcup_{j=1}^{m-1}[\mathcal{B}(b_j^2) \cup \mathcal{Y}(y_j^2)] \tag{3.3}$$

, in which b_i^1 and y_i^1 are ions generated from P_1, b_j^2 and y_j^2 are those ions came from P_2

3.2.3 Problem Definition

Because of the fact that the mass values acquired from mass spectrometers are not accurate, we use $\delta > 0$ to represent the maximum error bound of the mass spectrometers. Moreover, for spectrum S, we denote $\bar{S} = \{(x_i, h_i) \in \mathcal{M} | \exists y \in S, \text{ s.t. } |y - x_i| \leq \delta\}$, in which $\bar{S}$ is the subset of $\mathcal{M}$ containing all the peaks *explained* by the mass values in S.[2]

[2]The term *explained* means that a theoretical mass value matches with a peak in the acquired spectrum.

Let $S(P_1, P_2)$ be all the possible ion masses of P_1 and P_2, $S(P_1, P_2)$ can be computed by formula 3.3. Then $\overline{S(P_1, P_2)}$ contains all the peaks in $\mathcal{M}$ that can be explained by the ions of P_1 and P_2. Intuitively, the more and higher peaks included in $\overline{S(P_1, P_2)}$ indicates the more likely that $\mathcal{M}$ is generated from P_1 and P_2. Thus, the Mixture Spectra De Novo Sequencing problem can be formulated as follows: Given a mixture spectrum $\mathcal{M}$, and two precursor mass values MW_1 and MW_2, $|MW_1 - MW_2| \leq \Delta$ and a predefined error bound δ, we want to construct two peptides P_1 and P_2, such that $|||P_1|| + 20.03 - MW_1| \leq \delta$, $|||P_2|| + 20.03 - MW_2| \leq \delta$, and the following summation is maximized:

$$H(\overline{S(P_1, P_2)}) = \sum_{(x,h)\in\overline{S(P_1,P_2)}} h \tag{3.4}$$

The equation above is the summation of all the intensity values of the included peaks. It's worthy to notice that the scoring function mentioned above is totally replaceable. In practice, we can apply a more sophisticated scoring function which involves more influential factors than just the height of peaks.

3.3 Algorithms and Complexity

3.3.1 Formulation of Idea

As mentioned above, there will be new complications when dealing with mixture spectra. Firstly, the difficulty to design an appropriate scoring function to evaluate the similarity between the mixture spectra and the constructed candidate peptide pairs. The scoring function in Equation 3.4 considers the peak intensity of ions from both peptides, providing us an primitive way for such a task. Secondly, the difficulty to establish an efficient method to address the overlapping peaks. Overlapping peaks occur more frequently for co-sequenced peptides, which makes the interpretation more complicated and less accurate. Thirdly, the difficulty to establish an efficient mechanism to construct a pair of peptides simultaneously. By avoiding

the traditional idea of identifying multiple peptides iteratively, we explicitly model and intend to construct the co-sequenced peptides concurrently. Our proposed algorithm below can handle those difficulties effectively by gradually constructing prefix-suffix pairs for both peptides in a specially designated pathway, and when facing with the overlapping peaks, we only count each peak once in the scoring function.

Assume that $MW_1 = ||P_1|| + 20$ and $MW_2 = ||P_2|| + 20$, and let $A = a_1a_2...a_k$ be a prefix(N-terminus) of peptide $P_1 = a_1a_2...a_n$, the mass of the $b-$ion produced by a cleavage between a_i and a_{i+1} is represented as $||a_1a_2...a_i||_b = \Sigma_{1\leq x\leq i}||a_x|| + 1$, thus the mass of the $y-$ion from the same cleavage site is $MW_1 - ||a_1a_2...a_i||_b$. Let $B = a_na_{n-1}...a_{n-(t-1)}$ be the *reverse* string of a suffix(C-terminus) of peptide P_1, and we denote the mass of the $y-$ion by a cut between a_{n-i} and $a_{n-(i+1)}$ as $||a_na_{n-1}...a_{n-i}||_y = \Sigma_{n-i\leq x\leq n}||a_x|| + 19$, thus the mass of the $b-$ion from the same cut site is $MW_1 - ||a_na_{n-1}...a_{n-i}||_y$. Let $C = b_1b_2...b_u$ and $D = b_mb_{m-1}...b_{m-(v-1)}$ be a prefix and a *reverse* suffix of peptide $P_2 = b_1b_2...b_m$ respectively.

In order to simplify the symbolization in the next part, for a given string S, we use S^i to denote the substring of S from its leftmost amino acid to its i^{th} amino acid. For instance, $A^i = a_1a_2...a_i$ represents a fraction of A with i letters starting from the left terminus, on the other hand, $B^i = a_na_{n-1}...a_{n-(i-1)}$ represents a fraction of B which contains the leftmost i consecutive letters. In addition, we use $S_N(A)$ to denote the values of all the ions caused from every possible cleavage in A, then we have

$$S_N(A) = \bigcup_{i=1}^{k}[\mathcal{B}(||A^i||_b) \cup \mathcal{Y}(MW_1 - ||A^i||_b)]$$

Similarly, we use $S_C(B)$ to denote all the mass values induced by each different cut in the *reverse* suffix string B, then we will have the following equation:

$$S_C(B) = \bigcup_{i=1}^{t}[\mathcal{Y}(||B^i||_y) \cup \mathcal{B}(MW_1 - ||B^i||_y)]$$

Accordingly, for peptide P_2 and its prefix string C and *reverse* suffix D, we will have the

equations listed below:

$$S_N(C) = \bigcup_{i=1}^{u} [\mathcal{B}(\|C^i\|_b) \cup \mathcal{Y}(MW_2 - \|C^i\|_b)]$$

and

$$S_C(D) = \bigcup_{i=1}^{v} [\mathcal{Y}(\|D^i\|_y) \cup \mathcal{B}(MW_2 - \|D^i\|_y)]$$

Based on the four equations above, if it satisfies that $P_1 = A\alpha\overleftarrow{B}$, and $P_2 = C\beta\overleftarrow{D}$, $\alpha \in \Sigma, \beta \in \Sigma$, as shown in Figure 3.2, where $\overleftarrow{B}$ and $\overleftarrow{D}$ are the *reverse* strings of B and D, then the following equation is easy to obtain:

$$S(P_1, P_2) = S_N(A) \cup S_C(B) \cup S_N(C) \cup S_C(D) \tag{3.5}$$

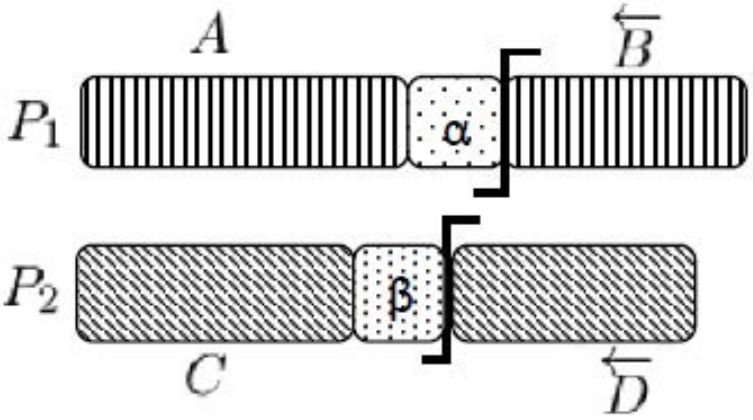

Figure 3.2: Exemplary diagram for peptides P_1 , P_2 and their prefix-suffix pairs.

This formula indicates that the Mixture Spectra De Novo Sequencing can be achieved by gradually constructing appropriate prefixes and suffixes of P_1 and P_2. In the following, we will describe a method that constructs the appropriate prefix-suffix pairs for both P_1 and P_2.

3.3.2 Algorithm for Candidate Computation

Given an amino acid string $s = s_1 s_2..s_n$, we call $||s|| = \Sigma_1^n ||s_i||$ the weight of s, and we use $s_> = s_1 s_2...s_{n-1}$ to denote the string with the the last amino acid removed. We assume $MW = \min\{MW_1, MW_2\}$ in the following research. For four strings A, B, C, D, we additionally define three useful sets $\mathcal{Q} = \{A, B, C, D\}$, $\mathcal{R} = \{A, B\}$, and $\mathcal{T} = \{C, D\}$. For simplicity, we denote $||A|| = ||A||_b$, $||B|| = ||B||_y$, $||C|| = ||C||_b$ and $||D|| = ||D||_y$ in the following section.[3]

Definition 3.1 *A pair of strings $\mathcal{E}$ and $\mathcal{F}$ are called Peers, if either of the following inequations holds: $||\mathcal{F}_>|| < ||\mathcal{E}|| \leq ||\mathcal{F}||$ or $||\mathcal{E}_>|| \leq ||\mathcal{F}|| < ||\mathcal{E}||$.*

Definition 3.2 *The string quadruple $\mathcal{Q} = \{A, B, C, D\}$ is called a Rigid Quartet if it satisfies the following conditions: (1) $||A|| + ||B|| \leq MW_1$ and $||C|| + ||D|| \leq MW_2$. (2) For any pair of strings $\mathcal{E}, \mathcal{F} \in Q$, $\mathcal{E}$ and $\mathcal{F}$ are Peers. (3) For any string $\mathcal{G} \in Q$, the following inequation holds: $||\mathcal{G}|| < MW - ||\mathcal{G}_>|| - 54$.*

Definition 3.3 *The string quadruple $\mathcal{Q} = \{A, B, C, D\}$ is called a General Quartet if it satisfies the following conditions: (1) $||A|| + ||B|| \leq MW_1$ and $||C|| + ||D|| \leq MW_2$. (2) A and B are Peers, and C and D are Peers.(3) For any string $\mathcal{G} \in Q$, the following inequation holds: $||\mathcal{G}|| < MW - ||\mathcal{G}_>|| - 54$.*

In Definition 3.2, the constraint of either $||\mathcal{F}_>|| < ||\mathcal{E}|| \leq ||\mathcal{F}||$ or $||\mathcal{E}_>|| \leq ||\mathcal{F}|| < ||\mathcal{E}||$ holding guarantees that the four elements $\{A, B, C, D\}$ we are constructing have always relatively close weight/mass values. The weight difference between any two elements are smaller than the

[3]Notice that $||A||_b$ represents the mass value of the corresponding $b-$ion, it equals to the summation of mass of amino acid residues and the extra mass of terminus. We denote $||A|| = ||A||_b$ here for the purpose of simplicity. It is same reason for denoting $||B|| = ||B||_y$, $||C|| = ||C||_b$ and $||D|| = ||D||_y$.

weight of an amino acid. When satisfying the constraint, we can see that every pair of element strings in the quadruple are indeed *Peers* in Definition 3.1. The inequation $\|\mathcal{G}\| < MW - \|\mathcal{G}_{>}\| - 54$ is also a necessary constraint for the proposed algorithm in the following part. We use an example to argue for this. Suppose that we have a suffix string B, and want to add another amino acid/letter to it.

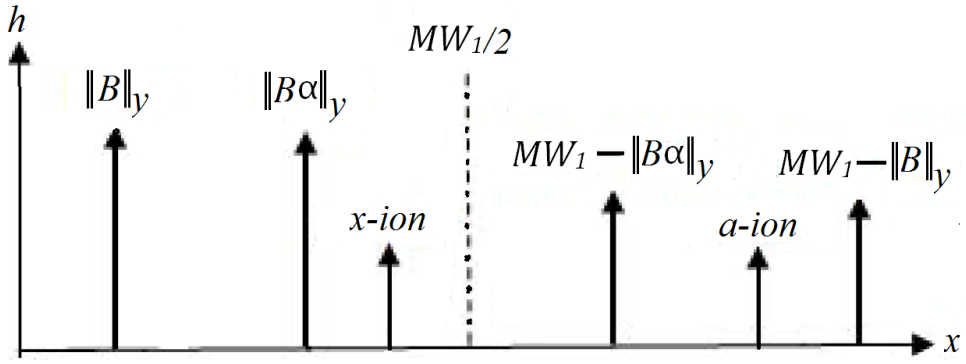

Figure 3.3: The inequation constraint in Definition 1. The mass of the $x-ion$ is $\|B\alpha\|_y + 26$, and the mass of the $a-ion$ is $MW_1 - \|B\|_y - 28$.

We know that B is from the C-terminus, its weight is calculated as $\|B\|_y$, the corresponding $b-$ion generated from the same cleavage site is $MW_1 - \|B\|_y$. If we add letter α to B, theoretically there will be a new $y-$ion and a new $b-$ion in the spectrum, whose mass values are $\|B\alpha\|_y$ and $MW_1 - \|B\alpha\|_y$. In the computation, we require that at least the inequation $\|B\alpha\|_y + 26 < MW_1 - \|B\|_y - 28$ holds. As showed in Figure 3.3, it indicates that after we add amino acid α, the new $y-$ion $\|B\alpha\|_y$ will possibly have overlapping peaks with the new $b-$ion $MW_1 - \|B\alpha\|_y$, but will never have overlapping peaks with the former $b-$ion $MW_1 - \|B\|_y$.

Corollary 3.1 *If $Q = \{A, B, C, D\}$ is a Rigid Quartet, then it is also a General Quartet.*

Proof According to Definition 3.2 and Definition 3.3, we see the difference in the definition of Rigid Quartet and General Quartet is only the second constraint.

If a string quadruple $Q = \{A, B, C, D\}$ is a Rigid Quartet, it indicates that any two strings of Q are *Peers*. Meanwhile, if a string quadruple $Q = \{A, B, C, D\}$ is a General Quartet, it means that A and B are *Peers*, at the same time C and D are also *Peers*. It is easy to conclude that a Rigid Quartet is also a General Quartet. □

Lemma 3.1 *Let $Q = \{A, B, C, D\}$ be a Rigid Quartet and let $a \in \Sigma$ be an amino acid. If A has the smallest weight among the quadruple and $||A|| + ||a|| + ||B|| \leq MW_1$ and $||A|| + ||a|| < MW - ||A|| - 54$, then the quadruple $\{Aa, B, C, D\}$ is also a Rigid Quartet.*

Proof We have the following inequations hold: $||B_>|| \leq ||A|| \leq ||B||$, $||C_>|| \leq ||A|| \leq ||C||$, and $||D_>|| \leq ||A|| \leq ||D||$. We know that $||Aa|| > ||A||$, so if $||Aa|| \leq ||B||$, then we have $||B_>|| < ||Aa|| \leq ||B||$. If $||Aa|| > ||B||$, then we have $||A|| \leq ||B|| < ||Aa||$. So we know that the string pair $\{Aa, B\}$ will always satisfy the restrictions in Definition 3.2.

Accordingly, we can easily prove that string pairs $\{Aa, C\}$ and $\{Aa, D\}$ also satisfies the restrictions in Definition 3.2, and we already have$||A|| + ||a|| + ||B|| \leq MW_1$. Together with the precondition $||A|| + ||a|| < MW - ||A|| - 54$, we will have the conclusion that $Q(Aa, B, C, D)$ is a Rigid Quartet. □

It also works for the case that B, C or D is the smallest. Lemma 3.1 tells that for a given Rigid Quartet, if we extend the smallest weighted string with one amino acid, the new quadruple we get is also a Rigid Quartet.

Lemma 3.2 *Let $Q(A, B, C, D)$ be a Rigid Quartet, and $Q(Aa, B, C, D)$, $a \in \Sigma$ also be a Rigid Quartet, then A has the smallest weight among the quadruple.*

Proof If $||A||$ is not the smallest weight, without losing generality, we assume $||A|| > ||B||$. Because $\{A, B, C, D\}$ is a Rigid Quartet, then the following inequation holds: $||B|| < ||A|| <$

$||Aa||$. it means for the string pair $\{Aa, B\}$, neither of the inequations $||B_{>}|| < ||A|| \leq ||B||$ and $||A_{>}|| \leq ||B|| < ||A||$ holds, then $Q(Aa, B, C, D)$ is not a Rigid Quartet. This is a contradiction with the condition.

Similarly, we can prove that if $||A|| > ||C||$ or if $||A|| > ||D||$, $Q(Aa, B, C, D)$ will not satisfy the constraints of Rigid Quartet. □

Lemma 3.2 indicates that a Rigid Quartet comes from a former Rigid Quartet by extending the smallest weighted string with one more amino acid.

Each of the four elements in a $Q = \{A, B, C, D\}$ represents an amino acid sequence or a fraction of a peptide sequence. Because A and C are the prefixes retained on the N-terminus of the corresponding peptides, in computation they have initial weight/mass values. In the initialization, we assume that the prefix is an empty sequence in which its mass value only contains the mass of the N-terminus(mass of a proton). From Figure 3.1 we know after initialization $||A_0|| = 1$ and $||C_0|| = 1$. Accordingly, B and D are the suffixes retained on the C-terminus of the corresponding peptides, they also have initial mass/weight values. The initial mass of a suffix containing empty sequence is the mass of the C-terminus plus the attached protons, based on Figure 3.1 we know $||B_0||^0 = 19$ and $||D_0||^0 = 19$.

Remark The quadruple $Q = \{A_0, B_0, C_0, D_0)\}$ in which each of the four elements are empty sequence containing only initial mass value is a Rigid Quartet.

Lemma 3.3 *Let $Q(A, B, C, D)$ be a General Quartet, and letters $a \in \Sigma$, $b \in \Sigma$, let $||A||$ be the smaller weighted one in $\mathcal{R}$, and $||C||$ be the smaller weighted one in $\mathcal{T}$. If $||A||+||a||+||B|| \leq MW_1$, $||C|| + ||b|| + ||D|| \leq MW_2$, $||A|| + ||a|| < MW - ||A|| - 54$ and $||C|| + ||b|| < MW - ||C|| - 54$, then both $Q(Aa, B, C, D)$ and $Q(A, B, Cb, D)$ are General Quartet(s).*

Proof Because $Q(A, B, C, D)$ is a General Quartet and $||A|| < ||B||$, then we have the following inequation holds: $||B_>|| \leq ||A|| \leq ||B||$. We know that $||Aa|| > ||A||$, so if $||Aa|| \leq ||B||$, then we have $||B_>|| < ||Aa|| \leq ||B||$. If $||Aa|| > ||B||$, then we have $||A|| \leq ||B|| < ||Aa||$. In either case, the two elements A and B are *Peers* which will satisfy the second constraint of the Definition 3.3.

We already have the conditions $||A|| + ||a|| + ||B|| \leq MW_1$ and $||A|| + ||a|| < MW - ||A|| - 54$ hold, therefore all the tree constraints of Definition 3.3 are satisfied, which indicates the new quadruple $Q(Aa, B, C, D)$ is also a General Quartet.

Similar arguments can be applied to prove that the new quadruple $Q(A, B, Cb, D)$ is also a General Quartet. □

Similar to Lemma 3.1, Lemma 3.3 tells that if we add a amino acid to the smaller element in the General Quartet, the newly constructed quadruple is also a General Quartet.

Lemma 3.4 *Let $Q(A, B, C, D)$ and $Q(Aa, B, C, D)$ both be Rigid Quartets, and $|MW_1 - MW_2| \leq \Delta$. Denote $\Lambda = \overline{S_N(A_>) \cup S_C(B_>) \cup S_N(C_>) \cup S_C(D_>)}$, then we have:*

$$(1)\ \overline{\mathcal{B}(||Aa||)} \cap \Lambda = \phi$$

and

$$(2)\ \overline{\mathcal{Y}(MW_1 - ||Aa||)} \cap \Lambda = \phi$$

Proof From Lemma 3.2. We know that $||A|| \leq ||B||$, $||A|| \leq ||C||$, and $||A|| \leq ||D||$. Without losing generality, we assume $MW = MW_1$ and $MW_2 = MW_1 + \Delta$.

Let Z be any prefix of A, and $Z \neq A$. We know that $\mathcal{B}(||Aa||)$ is apart from $\mathcal{B}(||Z||)$. Indeed we have $[\mathcal{B}(||Aa||) - \mathcal{B}(||Z||)]_{\min} = 2 \times \min_{a \in \Sigma}||a|| - (28 + 17) > 69$. This means that $\overline{\mathcal{B}(||Aa||)} \cap \overline{\mathcal{B}(||Z||)} = \phi$. We have $||Aa|| < MW - ||A|| - 54 \leq MW_1 - ||Z|| - (54 + \min ||a||_{a \in \Sigma})$.

this means that $\overline{\mathcal{B}(||Aa||)} \cap \overline{\mathcal{Y}(MW_1 - ||Z||)} = \phi$. Therefore, we have the following conclusion: $\overline{\mathcal{B}(||Aa||)} \cap \overline{S_N(A_>)} = \phi$.

Similarly, assume Z be any prefix of B, and $Z \neq B$, we have the following inequations: $||Aa|| \geq ||A|| + \min ||a||_{a\in\Sigma} \geq ||B_>|| + \min ||a||_{a\in\Sigma} \geq ||Z|| + \min ||a||_{a\in\Sigma}$, this means that $\overline{\mathcal{B}(||Aa||)} \cap \overline{\mathcal{Y}(||Z||)} = \phi$. We also know that $||Aa|| < MW - ||A|| - 54 \leq MW_1 - ||B_>|| - 54$, this indicates that $\overline{\mathcal{B}(||Aa||)} \cap \overline{\mathcal{B}(MW_1 - ||Z||)} = \phi$. Therefore we have the following equation: $\overline{\mathcal{B}(||Aa||)} \cap \overline{S_C(B_>)} = \phi$.

Similarly, assume Z be any prefix of C, and $Z \neq C$, we have the following holds: $||Aa|| \geq ||C_>|| + \min ||a||_{a\in\Sigma} \geq ||Z|| + \min ||a||_{a\in\Sigma}$. This tells that $\overline{\mathcal{B}(||Aa||)} \cap \overline{\mathcal{B}(||Z||)} = \phi$. We have $||Aa|| < MW - ||A|| - 54 \leq MW_2 - ||Z|| - (54 + \Delta)$. This inequation indicates that $||Aa||$ is apart from $||MW_2|| - ||Z||$, $\overline{\mathcal{B}(||Aa||)} \cap \overline{\mathcal{B}(MW_2 - ||Z||)} = \phi$. Therefore we have $\overline{\mathcal{B}(||Aa||)} \cap \overline{S_N(C_>)} = \phi$.

Similarly we can prove that $\overline{\mathcal{B}(||Aa||)} \cap \overline{S_C(D_>)} = \phi$. Therefore (1) is proved.

Similar arguments can be applied to $\overline{\mathcal{Y}(MW_1 - ||Aa||)}$ to prove (2). □

Lemma 3.4 shows that in the step of extending a Rigid Quartet $Q(A, B, C, D)$ by adding an amino acid to its smallest elements, a new Rigid Quartet $Q(Aa, B, C, D)$ is constructed, and all the new mass values generated because of this extension behaviour are contained in the following two sets: $\overline{\mathcal{B}(||Aa||)}$ and $\overline{\mathcal{Y}(MW_1 - ||Aa||)}$, and these two sets can only intersect with the mass values in $\mathcal{B}(||A||)$, $\mathcal{B}(||C||)$, $\mathcal{Y}(||B||)$, $\mathcal{Y}(||D||)$, $\mathcal{Y}(MW_1 - ||A||)$, $\mathcal{Y}(MW_2 - ||C||)$, $\mathcal{B}(MW_1 - ||B||)$ and $\mathcal{B}(MW_2 - ||D||)$. It means that in computation, when extending a Rigid Quartet, all the new peaks we have to count are in $\overline{\mathcal{B}(||Aa||)}$ and $\overline{\mathcal{Y}(MW_1 - ||Aa||)}$, and we also have to exclude the possible overlapping peaks that are already included in $\mathcal{B}(||A||)$, $\mathcal{B}(||C||)$, $\mathcal{Y}(||B||)$, $\mathcal{Y}(||D||)$, $\mathcal{Y}(MW_1 - ||A||)$, $\mathcal{Y}(MW_2 - ||B||)$, $\mathcal{B}(MW_1 - ||C||)$ and $\mathcal{B}(MW_2 - ||D||)$.

Lemma 3.5 *Let $\mathcal{Q}(A, B, C, D)$ be a Rigid Quartet, and a letter $a \in \Sigma$. We denote $f_v = \mathcal{B}(v) \cup \mathcal{Y}(MW_1 - v)$, $f_w = \mathcal{Y}(w) \cup \mathcal{B}(MW_1 - w)$, $f_m = \mathcal{B}(m) \cup \mathcal{Y}(MW_2 - m)$ and $f_n = \mathcal{Y}(n) \cup \mathcal{B}(MW_2 - n)$, and we define function $\Psi(v, w, m, n) = \overline{f_v \cup f_w \cup f_m \cup f_n}$ and let $\Gamma = \overline{S_N(A) \cup S_C(B) \cup S_N(C) \cup S_C(D)}$.*

(1) If $\mathcal{Q}(Aa, B, C, D)$ is a Rigid Quartet, and $f_1(u, v, w, m, n) = H(\overline{\mathcal{B}(u) \cup \mathcal{Y}(MW_1 - u)} \backslash \Psi)$, in which H is defined in Formula 3.4, then

$$H(\overline{S_N(Aa) \cup S_C(B) \cup S_N(C) \cup S_C(D)})$$
$$= f_1(||Aa||, ||A||, ||B||, ||C||, ||D||) + H(\Gamma).$$

(2) If $\mathcal{Q}(A, Ba, C, D)$ is a Rigid Quartet, and $f_2(u, v, w, m, n) = H(\overline{\mathcal{B}(MW_1 - u) \cup \mathcal{Y}(u)} \backslash \Psi)$, in which H is defined in Formula 3.4, then

$$H(\overline{S_N(A) \cup S_C(Ba) \cup S_N(C) \cup S_C(D)})$$
$$= f_2(||Ba||, ||A||, ||B||, ||C||, ||D||) + H(\Gamma).$$

(3) If $\mathcal{Q}(A, B, Ca, D)$ is a Rigid Quartet, and $f_3(u, v, w, m, n) = H(\overline{\mathcal{B}(u) \cup \mathcal{Y}(MW_2 - u)} \backslash \Psi)$, in which H is defined in Formula 3.4, then

$$H(\overline{S_N(A) \cup S_C(B) \cup S_N(Ca) \cup S_C(D)})$$
$$= f_3(||Ca||, ||A||, ||B||, ||C||, ||D||) + H(\Gamma).$$

(4) If $\mathcal{Q}(A, B, C, Da)$ is a Rigid Quartet, and $f_4(u, v, w, m, n) = H(\overline{\mathcal{B}(MW_2 - u) \cup \mathcal{Y}(u)} \backslash \Psi)$, in which H is defined in Formula 3.4, then

$$H(\overline{S_N(A) \cup S_C(B) \cup S_N(C) \cup S_C(Da)})$$
$$= f_4(||Da||, ||A||, ||B||, ||C||, ||D||) + H(\Gamma).$$

Proof (1) Let $u = ||Aa||$, $v = ||A||$, $w = ||B||$, $m = ||C||$, and $n = ||D||$ respectively, and we denote $\Lambda = \overline{S_N(A_>) \cup S_C(B_>) \cup S_N(C_>) \cup S_C(D_>)}$. We know the fact that $\Psi(v, w, m, n) \subset \Gamma$. Therefore

we have the following deductions:

$$\begin{aligned}
&\overline{S_N(Aa) \cup S_C(B) \cup S_N(C) \cup S_C(D)} \\
&= \Gamma \cup \overline{B(u) \cup Y(MW_1 - u)} \\
&= \Gamma \cup (\overline{B(u) \cup Y(MW_1 - u)} \backslash \Psi(v, w, m, n))
\end{aligned}$$

According to Lemma 3.4, we have the following equations:

$$\begin{aligned}
&\Gamma \cap (\overline{B(u) \cup Y(MW_1 - u)} \backslash \Psi(v, w, m, n)) \\
&= \overline{B(u) \cup Y(MW_1 - u)} \cap (\Gamma \backslash \Psi(v, w, m, n)) \\
&\subset \overline{B(u) \cup Y(MW_1 - u)} \cap \Lambda \\
&= \phi
\end{aligned}$$

Therefore we have the equations hold that proves (1).

$$\begin{aligned}
&H(\overline{S_N(Aa) \cup S_C(B) \cup S_N(C) \cup S_C(D)}) \\
&= H(\Gamma) + f_1(u, v, w, m, n)
\end{aligned}$$

Proof of (2), (3), (4) is very similar to those above, therefore omitted. □

Lemma 5 indicates that the set for a newly generated quadruple which contains all the matched peaks can be calculated from the previous quadruple by adding some new peaks contained in f. The function H in Formula 3.4 is the summation of all the explained peaks contained in this new set.

For string pair $\mathcal{R} = (A, B)$, and string pair $\mathcal{T} = (C, D)$, we define three relations between $\mathcal{R}$ and $\mathcal{T}$: *Neighbouring, Crossing, and Nesting*. Assuming that $||A|| \leq ||B||$, and $||C|| \leq ||D||$.

- If $||A|| \leq ||B|| \leq ||C|| \leq ||D||$, then we say the two string pairs $\mathcal{R}$ and $\mathcal{T}$ are *Neighbouring*.
- If $||A|| \leq ||C|| \leq ||B|| \leq ||D||$, then we say the two string pairs $\mathcal{R}$ and $\mathcal{T}$ are *Crossing*.

- If $||A|| \leq ||C|| \leq ||D|| \leq ||B||$, then we say the two string pairs $\mathcal{R}$ and $\mathcal{T}$ are *Nesting*.

Suppose that for a *Rigid Quartet* $\mathcal{Q} = \{A, B, C, D\}$ in which $||A||$ is the smallest weighted. When adding letter $a \in \Sigma$, if it satisfies that $||A|| + ||a|| < MW - ||A|| - 54$, and $||A|| + ||a|| + ||B|| < MW_1$, then the newly constructed quadruple is also a *Rigid Quartet*, the algorithm proceeds to the next step. Meanwhile, if it satisfies that $||A|| + ||a|| < MW - ||A|| - 54$, and $||A|| + ||a|| + ||B|| = MW_1$, then the first peptide is already fully constructed. Moreover, if it satisfies $||A|| + ||a|| + ||B|| < MW_1$, and $||A|| + ||a|| \geq MW - ||A|| - 54$. In this case, we will have the following conclusion: $2 \times ||A|| + ||a|| \geq MW - 54$ and $2 \times ||B|| + ||a|| \geq MW - 54$, together we have $||A|| + ||B|| + ||a|| \geq MW - 54 \geq MW_1 - \Delta - 54$. It tells that if we add letter a to string A, the remaining weight will not fit in any other letter. We only consider the case that for some letter $b \in \Sigma$, that $||A|| + ||B|| + ||b|| = MW_1$ exactly. After we get one complete peptide sequence, we will continue to construct the other peptide. Lemma 3.6 below illustrates the details of the extension of the other peptide.

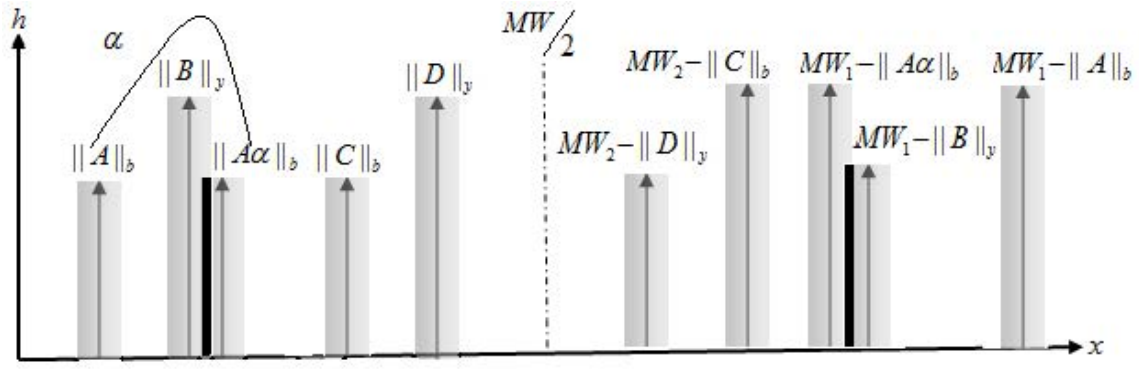

Figure 3.4: Illustration for growing a *Rigid Quartet*: $Q(A, B, C, D)$ is a *Rigid Quartet* in which $||A||$ is the smallest weighted, after adding one amino acid to A, the new quadruple $Q(Aa, B, C, D)$ is also a *Rigid Quartet*. In each round, we always try to extend the smallest weighted element. The shadow area surrounding each peak p refers to the distribution of peaks $\mathcal{B}(p)$ or $\mathcal{Y}(p)$. The blackened area indicates that there might be overlapping peaks located here.

Lemma 3.6 *Let $\mathcal{Q} = (A, B, C, D)$ be a Rigid Quartet, and let $||A||$ be the smaller weighted one*

in $\mathcal{R}$, and $||C||$ be the smaller weighted one in $\mathcal{T}$, and $||A|| \leq ||C||$. Denote

$$\Lambda = \overline{S_N(A_>) \cup S_C(B_>) \cup S_N(C_>) \cup S_C(D_>)}$$

and assume $a \in \Sigma$ is the letter to be added to $\mathcal{T}$. If for an amino acid $b \in \Sigma$, the following condition holds: $||A|| + ||b|| + ||B|| = MW_1$, then we have:

(1) If $\mathcal{R}$ and $\mathcal{T}$ are neighbouring or crossing, then $\mathcal{T}$ can be extended at most 2 times and we have [4]

$$\overline{\mathcal{B}(||Ca||) \cup \mathcal{Y}(MW_2 - ||Ca||)} \cap \Lambda = \phi$$

.

(2) If $\mathcal{R}$ and $\mathcal{T}$ are nesting, then $\mathcal{T}$ can be extended at most 5 times, and we have

$$\overline{\mathcal{B}(||Ca||) \cup \mathcal{Y}(MW_2 - ||Ca||)} \cap \Lambda = \phi$$

.

Proof (1) For the case of *Neighbouring*, we know that $||A|| \leq ||B|| \leq ||C|| \leq ||D||$ and we also have $||A|| + ||a_1|| + ||B|| = MW_1$. Without losing generality, we assume $MW = MW_1$ and $MW_2 = MW_1 + \Delta$. Then we know that $||C|| + ||D|| + ||a_1|| \geq ||A|| + ||B|| + ||a_1|| = MW_1 = MW_2 - \Delta$. From this inequation, we can infer $MW_2 - (||C|| + ||D||) \leq ||a_1|| + \Delta \leq \max ||a||_{a\in\Sigma} + \Delta$. Thus, string pair $\mathcal{T}$ can be at most extended $\lfloor(\max ||a||_{a\in\Sigma} + \Delta)/\min ||a||_{a\in\Sigma}\rfloor - 1 = 2$ times. The proof for both $\overline{\mathcal{B}(||Ca||)} \cap \Lambda = \phi$ and $\overline{\mathcal{Y}(MW_2 - ||Ca||)} \cap \Lambda = \phi$ is similar to Lemma 3.4.

For the case of *Crossing*, the proof is similar and therefore omitted here.

(2) From the relation of $\mathcal{R}$ and $\mathcal{T}$, we know that $||C|| + ||D|| \geq 2 \times ||A||$. Meanwhile we know that $||A|| + ||a_1|| + ||B|| = MW_1$ and $||B|| \leq ||A|| + \max ||a||_{a\in\Sigma}$. Together, we have $2 \times ||A|| + \max ||a||_{a\in\Sigma} +$

[4] We say $\mathcal{R}$ or $\mathcal{T}$ is *extended*, if we add a letter $a \in \Sigma$ to the smaller weighted string in $\mathcal{R}$ or $\mathcal{T}$, the new quadruple is at least a *General Quartet*.

$||a_1|| \geq MW_2 - \Delta$. It means $MW_2 - (||C|| + ||D||) \leq \max ||a||_{a\in\Sigma} + ||a_1|| + \Delta \leq 2 \times \max ||a||_{a\in\Sigma} + \Delta$, then we know that for $\mathcal{T}$, it can be extended at most $\lfloor (2 \times \max ||a||_{a\in\Sigma} + \Delta)/ \min ||a||_{a\in\Sigma} \rfloor - 1 = 5$ times.

It is easy to get that $\mathcal{B}(||Ca||) \cap S_N(A_>) = \phi$ and $\mathcal{B}(||Ca||) \cap S_N(C_>) = \phi$ and $\mathcal{B}(||Ca||) \cap S_C(D_>) = \phi$. From the conditions, we know $||A|| \leq ||C|| \leq ||D|| \leq ||B||$ and $||Ca|| < MW - ||C|| - 54$ and $||B_>|| \leq ||A||$, so we have $||Ca|| < MW_1 - ||A|| - 54 < MW_1 - ||B_>|| - 54$. We also have $||Ca|| \geq ||C|| + \min ||a||_{a\in\Sigma} \geq ||B_>|| + \min ||a||_{a\in\Sigma}$. Together, we get $\mathcal{B}(||Ca||) \cap S_C(B_>) = \phi$. Therefore, we prove $\overline{\mathcal{B}(||Ca||)} \cap \Lambda = \phi$. Similarly, we can prove $\overline{\mathcal{Y}(MW_2 - ||Ca||)} \cap \Lambda = \phi$. □

Our basic idea of Peptide De Novo Sequencing from Mixture Spectra is based on our mathematical model and proof above. As shown in Algorithm 1, we construct four strings $\{A, B, C, D\}$ following a specifically designated pathway. They are the N-terminus prefix and the reverse C-terminus suffix of the targeted peptide P_1 and P_2 respectively. The algorithm starts from growing a *Rigid Quartet*, according to Lemma 3.1, when we add one amino acid to the smallest weighted string, the newly acquired string quadruple is still a *Rigid Quartet*. The extending procedure is illustrated in Figure 3.4. In each step of extending a *Rigid Quartet*, the score for the string quadruple is calculated based on Lemma 3.5. Moreover, we can see from Figure 3.4 that when a peak is matched by two different ions, we are able to identify and locate the potential overlapping peaks in the computation procedure, and our basic guidelines for treating a overlapping peak is that we add the height of that peak only once. In every round of extending, an amino acid is always added to the smallest weighted sequence in the quadruple.

Usually in the last few steps, it will occur that the extending of the quadruple can't be constrained by the definition of *Rigid Quartet*. This happens when we've already found a candidate sequence for one peptide(or to say that the prefix-suffix pair for one peptide is ready to *merge* and the mass summation matches with the total molecular mass value). In such condition, we consider the quadruple as a *General Quartet* and continue to extend the second prefix-suffix

Algorithm 1 *De Novo* Sequencing of Mixture Spectra

INPUT: Given mixture spectrum $\mathcal{M}$, and two precursor mass values MW_1 and MW_2, $|MW_1 - MW_2| \leq \Delta$, and a predefined error bound δ, and the finest mass spectrometer calibration β

OUTPUT: Constructing two peptides P_1 and P_2, which satisfies $|||P_1|| + 20 - MW_1| \leq \delta$, $|||P_2|| + 20 - MW_2| \leq \delta$, and $H(P_1, P_2)$ is maximized.

1: Initializing all $DP[i, j, k, l] = -\infty$. $DP[]$ refers to a table in memory to store values calculated in the algorithm.
2: Let $MW = \min\{MW_1, MW_2\}$ and $DP[1, 19, 1, 19] = 0$
3: **for** Each possible combination of (x, y, m, n) **do**
4: **if** x is the smallest among (x, y, m, n) **then**
5: **for** $a \in \Sigma$ **do**
6: **if** $|x + y + ||a|| - MW_1| \leq \delta$ **then**
7: ProceedToEnd(x,y,m,n)
8: **else if** $x + ||a|| < MW - x - 54$ **then**
9: Let $loc \leftarrow (x + ||a||, y, m, n)$
10: $DP[loc] = \max \begin{cases} DP[loc] \\ DP[x, y, m, n] + f_1 \end{cases}$
11: **else if** y or m or n is the smallest **then**
12: Similar as above by updating the corresponding $DP[]$ score based on Lemma 3.5
13: Compute $DP[x, y, m, n]$ for all x, y, m, n and $a_1 \in \Sigma$ and $a_2 \in \Sigma$ satisfying $|x + y + ||a_1|| - MW_1| \leq \delta$ and $|m + n + ||a_2|| - MW_2| \leq \delta$
14: Backtrack and Output the best matched peptide pairs $Aa_1\overleftarrow{B}$ and $Ca_2\overleftarrow{D}$ or output a candidate pair list

pair until we obtain the whole sequence. At the same time, from Lemma 3.6, we know that when this situation happens, the following computation for extending the *General Quartet* can be completed in constant time, and we call this procedure ProceedToEnd in Algorithm 2. In the procedure of extending a *General Quartet*, the score calculation is the same as to extend a *Rigid Quartet*, which is listed in Lemma 3.5. After all the possible amino acid extension is considered, we then rank all the string quadruples that satisfy the given conditions and output the highest scored candidate pair in the end.

3.3.3 Complexity Analysis

We give the complexity of the proposed algorithm in the following Theorem 3.1.

Algorithm 2 ProceedToEnd: Continue to Construct the Second Peptide

ProceedToEnd: This procedure works in an recursive procedure to finish constructing the second peptide.

1: **procedure** ProceedToEnd(x, y, m, n)
2: **if** (x, y) completed, to extend (m, n) **then**
3: **if** $m \leq n$ **then**
4: **for** $a \in \Sigma$, s.t. $m + ||a|| < MW - m - 54$ **do**
5: Let $loc \leftarrow (x, y, m + ||a||, n)$
6: $DP[loc] = \max \begin{cases} DP[loc] \\ DP[x, y, m, n] + f_3 \end{cases}$
7: ProceedToEnd$(x, y, m + ||a||, n)$
8: **else**
9: **for** $a \in \Sigma$, s.t. $n + ||a|| < MW - n - 54$ **do**
10: Let $loc \leftarrow (x, y, m, n + ||a||)$
11: $DP[loc] = \max \begin{cases} DP[loc] \\ DP[x, y, m, n] + f_4 \end{cases}$
12: ProceedToEnd$(x, y, m, n + ||a||)$
13: **else if** (m, n) completed, to extend (x, y) **then**
14: Similar as above by updating the the corresponding $DP[]$ score based on Lemma 3.5

Theorem 3.1 *Algorithm 1 computes the optimal solution of the* Mixture Spectra De Novo Sequencing Problem *in time bounded by*

$$O((\frac{\max_{a\in\Sigma} ||a||}{\beta})^3 \times \frac{\delta}{\beta} \times |\Sigma| \times \frac{MW}{\beta})$$

Proof The procedure ProceedToEnd in Algorithm 2 is a recursive procedure. According to Lemma 3.6, we know the height of the recursive tree is bounded by a small number. In practice, we can use an extra data structure to build an index that contains all the possible amino acid compositions for a given weight. As shown in the proof of Lemma 6, the value is at most $(2 \times \max_{a\in\Sigma} ||a|| + \Delta)$. In this way the time efficiency for this procedure is even more improved. Lemma 3.5 tells us that the function $f(u, v, w, m, n)$ is to find the possible overlapping peaks in two sets and to calculate the summation of the height of all the explained peaks, it can be computed in $O(\frac{\delta}{\beta})$. In Algorithm 1, line 3 is a loop which enumerates all the possible combinations of (x, y, m, n) by step β. Based on Definition 1, we know that the four strings in the quadruple are bounded by each other. The mass difference between any two of them should

be smaller than the mass of an amino acid residue. Thus, the maximum times of steps in the enumeration is $(\frac{\max_{a\in\Sigma}\|a\|}{\beta})^3 \times \frac{MW}{\beta}$, and in each step algorithm has to try every character in the alphabet Σ. Overall, the time complexity is indeed linear to the mass of the peptide, MW here is the larger mass of P_1 and P_2.

3.4 Experiment Results and Discussion

Mixture spectra can occur quite frequently in a typical wet-lab mass spectrometry experiment. The ability to identify mixture spectra will be useful for improving the identification rate of the collected mass spectra. The *de novo* method for mixture spectra interpretation can be regarded as a supplementary method for peptide identification with mass spectrometry. It will find its application in assisting the traditional database search method or spectral library search method to identify mixture spectra. Because when considering two precursors in a single spectra, the possible candidate peptide pairs that fall into the required mass error bound is enormous [115]. In this case, an effective strategy for reducing the examination space is non-trivial, and a foregoing procedure of *de novo* sequencing is an useful operation in helping the filtration. In order to evaluate the effectiveness of the proposed algorithm, we use both simulated and real mixture spectra datasets to benchmark the performance.

3.4.1 Simulated Dataset

Due to the fact that large datasets of validated mixture spectra are not publicly available now, thus similar to those mentioned in [114] and [114], we created a dataset of 253 simulated mixture spectra to justify the performance of our algorithm. At the beginning, we selected some of the confidently identified MS/MS spectra with charge 2 from the PRIDE repository(Accession No. 17341-17350) [123]. And the two spectra we chose to merge have precursor m/z values

with difference less than 3 Th[5]. Precursor intensity has a crucial impact on MS/MS mixture spectra interpretation, and we want to inspect the influence of precursor abundances on the subsequent MS/MS spectra identification. Thus in the simulation, we use a coefficient to simulate the fact that two co-sequenced precursors may have different intensities. The linear combination of two spectra is adjusted by $\mathcal{M} = A + \alpha B$, in which α is the mixture coefficient. A and B are normalized based on the summation over every intensity value contained in the spectrum before merging, where α is used to rescale all the included peaks in B, and A always corresponds to the spectrum with high abundance. By tuning α, it is supposed that different fragment abundance levels can be simulated. Previous reports indicated that the abundance of fragments are indeed highly correlated to the abundance of the corresponding precursors, thus in our research, we use the coefficient α to emulate both precursor and fragments abundance. Furthermore, when merging two spectrum, we keep all the un-explained peaks to simulate the noise signals that exist universally in mass spectra. We compare a simple prototype of our algorithm with the notable commercialized software PEAKS [17] *de novo* online version on the simulated dataset to evaluate the effectiveness of the proposed algorithm. We use PEAKS 6.0 software to identify two peptides in an iterative manner, first identify the peptide with larger precursor intensity and then identify the second peptide from the spectra with the explained peaks of the first peptide subtracted. The experimental results are show in the following tables. We mainly focus on two meaningful aspects of the experimental results: the number of reported pairs in which both peptides have less than 4 incorrect characters, shown in Table 3.3, and the number of reported pairs that both peptides have longer than 3 consecutive correct characters, shown in Table 3.4.

From the Table 3.3 above, we can see that in the case that both precursors have comparable abundances, our algorithm gives more positive results. This is reasonable because our proposed algorithm formulates both peptides concurrently and consider the mixture spectra *de*

[5]The Th or $Thomson$ is a unit of mass-to-charge ratio that appears in the field of mass spectrometry.

Table 3.3: Number of reported pairs in which both peptides have less than 4 incorrect characters.

Coefficient α	Our Algorithm	PEAKS Denovo	Total
1.0	**100(39.5%)**	89(35.2%)	
0.8	**95(37.5%)**	94(37.1%)	(253)
0.6	89(35.2%)	106(41.9%)	

novo sequencing problem as a whole, therefore gives this method better performance under this circumstance. While for PEAKS working in an iterative manner, in case that two peptides with both relatively high intensities being fragmented together, the relatively higher peaks of one peptide will confuse the engine when constructing another peptide. However, in case that one precursor has apparently high intensity than another precursor, PEAKS *de novo* showed a vast improvement in identification accuracy. This is perhaps because when the precursor with higher intensity is reported and all its related fractions are removed, in the remaining spectrum, there will be much less interfering peaks from the first peptide, and the fraction peaks coming from the lower intensity precursor will be strong peaks, which is very helpful for the identification of the second peptide. A potential scenario for the application of our proposed algorithm is like this: When analysing the raw dataset collected from mass spectrometers, we look into the survey scans to check if there are more than one monoisotopic m/z values, because this highly indicates that the corresponding MS/MS spectra is a mixture spectra. It is supposed that this task can be done both manually and automatically by software. The objective is to locate the mixture spectra in the whole dataset and most importantly to find the mass and intensity values of the selected precursors. The abundance information of the co-sequenced precursors in the survey scan can help us to make decision whether to apply a traditional *de novo* sequencing method, or the proposed mixture spectra *de novo* method.

The number and ratio of reported pairs that both peptides have consecutive correct letters longer than 3 are shown in Table 3.4 above. The result shown in Table 3.4 above is consistent with that in Table 3.3, our algorithm works better for comparable abundance precursors. From

Table 3.4: Number of reported pairs that both peptides have consecutive correct letters longer than 3.

Coefficient α	Our Algorithm	PEAKS Denovo	Total
1.0	**172(68.0%)**	140(55.3%)	
0.8	**168(66.4%)**	153(60.5%)	(253)
0.6	161(63.6%)	164(64.8%)	

this table, we can see that the ratio that both peptides have tags longer than 3 is quite noticeable, which depicts another possible scenario of applying the proposed algorithm, to serve in the filtration process of using database search idea to identify mixture spectra. This application scenario relies on the traditional idea of integrating *de novo* sequencing results with database search method to accurately and sensitively identify MS/MS spectra.

3.4.2 Real Mixture Spectra

To further evaluate the efficiency of the proposed algorithm, we run the published software package MSPLIT [114] on a tryptic yeast dataset released on PRIDE repository to obtain some real mixture spectra. The yeast dataset is collected from an Ion Trap mass spectrometer. When analyzing a spectrum, MSPLIT will always identify some top-scoring peptide pair for any given query spectrum. To assess whether a match is significant, MSPLIT then applies an empirical strategy to classify the matched pair into three possible outcomes: No match, Single-peptide match, and Mixture match. To allow MSPLIT to run in the target/decoy strategy, we firstly downloaded a yeast spectral library from NIST(Ion Trap/2012.06.20 version), and then we apply the SpectraST [124] software to generate a concatenated target/decoy spectra library. By searching against this concatenated spectral library, MSPLIT can filter and report the identified mixture spectra under a given FDR. In this experiment, the number of MS/MS spectra analyzed by MSPLIT is 5244 in total, and MSPLIT reported 326 identifications in the final step, in which 24 are identified as mixture match, this result is to some extent consistent with the estimated ratio reported in [114] that mixture spectra consist of approximately 5%

of all identifiable spectra in the yeast dataset. In our experiment, we further filtered those reported mixture spectra according to our current requirements. We only select those spectra in which both precursors have charge 2, and contain no Post-translational Modifications due to the limitation of our current software prototype. The renowned database search software PEAKS DB [10] is used here to re-confirm those mixture spectra reported by MSPLIT. PEAKS DB is a popularly accepted software which matches each spectrum with one peptide sequence from protein database, and it can not automatically report multiple peptides from a mixture spectrum. Therefore, we use PEAKS DB in the following way for mixture spectra validation: each individual peptide is searched against the protein database with the other peptide sequence removed from the database to eliminate the potential interference between two peptides. Only those mixture spectra in which both peptides can be confirmed by PEAKS DB are kept. After all those steps, we obtained 7 distinct mixture spectra. Among those mixture spectra, our proposed algorithm can give useful[6] results for 6 of them. The reported results for the six mixture spectra are listed in Table3.5.

Table 3.5: Six sequence pairs with useful results reported by our algorithm.

m/z	DB Confirmed	Algorithm Given Pairs
553.321	GLILVGGYGTR	WP**LV**N**YGTR**
553.7795	GPPGVFEFEK	**GPPGV**HAASG**EK**
419.724	ASIASSFR	**ASLAS**HP**R**
420.2585	IAGLNPVR	**LAGL**AAAP**R**
506.78	FHLGNLGVR	**FH**ADPN**GVR**
507.2820	KFPVFYGR	**QFPV**NPV**GR**
600.2756	GYSTGYTGHTR	SSFS**GYTGHTR**
600.340	DAGTIAGLNVLR	NE**TLAGLNVLR**
462.706	DNEIDYR	**DNE**DL**YR**
463.3077	IVAALPTIK	VL**AALPTLK**
521.253	YSDFEKPR	**YSDF**PGSL**R**
521.7932	GAIAAAHYIR	Q**LAAAHYLR**

[6]When we say *useful*, it means that both reported peptides should have consecutive correct letters longer than 3.

In the reported results, the Isoleucine(I) and Leucine(L) are always treated as identical amino acids. High resolution/accuracy mass spectrometry data will surely provide great help for the computational analysis, but currently based on the accuracy of the proteomics data we use, in this experiment, the error tolerance for fragment peaks we chose is 0.2 Dalton. Though the Lysine(K) and Glutamine(Q) have slightly different residue mass values, they can not be separated with the mass tolerance configured in the current experiment, therefore they are also treated as the same. In Table 3.5, there are some interesting points worth noticing. Some incorrect results are due to identical mass values. For instance, in the first entry of Table 3.5, we noticed that the mass value of amino acid 'N' is indeed equal to that of 'GG'. And in the last entry, the mass value of 'Q' is actually equal to the summation of 'GA'. Some incorrect results are simply different permutations of the same group of letters. For instance, for both peptides in the fifth entry, the algorithm gave correct amino acid compositions, only a slight change in orders for a few letters. This is probably caused by the incomplete fragmentation of the selected precursors or the parameters configuration in the current experiment. Those kinds of partially correct or homologous results are very useful for assisting a subsequent database search to accurately identify peptides. MSPLIT is a library search method which can only report peptides contained in the spectra library. However, the size of the current spectral library is limited, which means for some spectra it may not report the best matched peptides. We use an additional database search process to guarantee that the collected spectra are indeed mixture spectra. Our proposed method has the potential to extend to more complex situations, for instance, co-fragmented precursors with different charge states, or precursor peptides containing PTMs.

Mixture spectra identification is a complicated, but promising research topic. The proteomics research community is at a stage of just realizing the significance of the existence of mixture spectra in tandem mass spectrometry, but still not offering an ideal solution yet. Moreover, the arising of DIA(Data Independent Acquisition) makes such computational approaches for analyzing mixture spectra more desired. In this chapter, we formulated the MIXTURE SPEC-

TRA DE NOVO SEQUENCING problem mathematically, and then proposed a dynamic programming algorithm for solving the problem. The performance evaluation of the algorithm on both simulated datasets and real mixture spectra demonstrated the meritorious aspects of our proposed algorithm. To provide better evaluation for how much help the proposed algorithm can actually offer is highly dependent on the public availability of confidently validated mixture dataset in the future. We will focus on seeking better scoring function for matching a pair of peptide sequences with a mixture spectra, and then we will try to develop database or spectral library methods by incorporating a *de novo* procedure to improve identification accuracy in the next phase. Another thing worth in-depth investigating is the evaluation of the survey scans to find if there are mixture spectra present, and the utilization of useful information lying behind the survey scans to assist the subsequent MS/MS spectra interpretation.

Chapter 4

De Novo Assisted Database Search

4.1 Introduction

Mixture spectra are observed quite frequently in mass spectrometry experiment which result from the concurrent fragmentation of multiple precursors, and the occurrence of mixture spectra increases as the complexity of the peptide mixture submitted to the mass spectrometer increases and it also depends on the width of the precursor ion selection window. Roughly, a mixture spectrum is defined as an MS/MS spectrum generated from the concurrent fragmentation of two or more peptides, which is theoretically modelled as the linear combination of two single-peptide spectra.

Although advances have been achieved in the past few years, the research on characterizing mixture spectra is still far from satisfactory. Despite of the rapid growth of publicly available spectral libraries, methods based on spectral library search will lose effectiveness when the target peptides have not been observed or identified before. Meanwhile, the research of *de novo* sequencing of mixture spectra is currently in its very primitive phase which suffers the problem of low accuracy in computation, more research will be necessary before it become practical for real application. For the large volumes of data generated in proteomics experi-

ments, sequence database searching combined with statistical validation of the search results is currently the only realistic method for assigning spectra to peptide sequences with relatively high confidence. Generally, database searching approach have in common that they attempt to find the best match between the acquired spectra and the spectra predicted from the peptide sequence in a protein database that have similar mass values with the precursor ion in the mass spectrometer.

Most of the mainstream database search engines assume that the acquired spectra originated from a single precursor, they therefore fail to identify the sequences of concurrently fragmented precursors, unless the spectra are dominated by fragment ions of one precursor, making the fragments of the other precursors appear as noise, and in this case only the dominant species can be correctly identified. Similar to the identification of single-peptide MS/MS spectra by comparison against all peptide entries in a database of know protein sequence, our general perspective is to compare mixture spectra against a pair of peptides acquired from a given protein database. However, when considering two precursors in a single spectra, the number of possible candidate peptides that fall into the required mass error bound is always enormous[115]. Efficient filtration strategy is highly desired to avoid the huge computational overload of searching the spectrum against all possible pairs in the protein database. Also, database search method requires effective scoring function to measure the quality of a match between a candidate and a given spectrum. Typically, the scoring function will give some reward for matches between observed peaks in the spectrum and theoretical masses from the candidate peptides, and as well impose penalties for unexplained spectrum peaks or missing theoretical ion masses from the candidate peptide. Although scoring a peptide against an MS/MS spectrum is a well studied problem in the research of mass spectrometry based proteomics, very few scoring functions are proposed to deal with the mixture spectra.

In this chapter, we will formulate the mixture spectra identification problem formally, and propose an approach for matching mixture MS/MS spectra with a pair of peptides from a

protein database by incorporating a special filtration strategy assisted with the preliminary *de novo* sequencing results. Also, in this research we introduced a method to estimate the mixture coefficient of the two co-sequenced peptides, which represents the relative abundance of the peptides when appear in the fragmentation cell. Experimental results demonstrated that when equipped with such filtration process, the correct matches can be found by only considering a minuscule fraction of all possible pairs.

4.2 Notations and Problem Formulation

4.2.1 Basic Notation

Similar to the notations in Chapter 3, we assume that a mixture spectrum $\mathcal{M}$ can be represented by a peak list $\mathcal{M} = \{(x_i, h_i)|i = 1, 2, ..., n\}$ in which each element (x_i, h_i) represents a peak in the spectrum. For each element (x_i, h_i) contained in the mixture spectrum $\mathcal{M}$, x_i is the m/z value when going through the mass analyzer and h_i is the intensity of the peak denoting the occurrence of the ions observed in the detector. Assume that associated with the given query mixture spectrum $\mathcal{M}$, we have two molecular weight MW_1 and MW_2 to denote the precursor mass values of the two peptides that satisfy $||MW_1 - MW_2|| \leq \Delta$, and Δ here is a small value predefined by the width of the mass spectrometer selection window. In current shotgun proteomics conducted with the Data-dependent Acquisition(DDA) strategy, the selection of the precursor ions are bounded by a very small isolation window usually just a few *Dalton*s wide which indicates that the molecular weights of the concurrently fragmented precursors are very similar to each other. A mixture spectrum is conventionally modelled as the linear combination of two spectra generated by the fragmentation of peptide P_1 and P_2 respectively. More formally, we formulated a mixture spectrum $\mathcal{M}$ as $\mathcal{M} = A + \alpha B$, where A and B are the MS/MS spectra generated by the fragmentation of two individual peptides P_1 and P_2 respectively, and the mixture coefficient α represents the relative abundance of A and B when being sequenced.

Without losing generality, we assume that both A and B are scaled to the same magnitude, and A always corresponds to the peptide with higher abundance level, therefore we have $0 \leq \alpha \leq 1$. In addition, we use Σ to denote the alphabet of 20 different types of amino acids. For an amino acid $a \in \Sigma$, we use $||a||$ to symbolize the mass of the amino acid residue. Let $P = a_1 a_2 ... a_k$ be the string of amino acids, we define the residue mass of the peptide as $||P|| = \Sigma_{1 \leq j \leq k} ||a_i||$ and the actual mass of the peptide as $||P|| + ||\mathrm{H_2O}||$.

4.2.2 Problem Formulation

Our primary idea for mixture spectra identification is to search the query mixture spectra against a protein database to find the best matched peptide pair. Thus, the Mixture Spectra Database Search Problem can be formulated as follows: Given a mixture spectrum $\mathcal{M}$, two precursor mass values MW_1 and MW_2, a predefined error bound δ, and a protein database D, we want to find a coefficient α and a pair of peptides P_1 and P_2 from D that maximize the matching score under a specific scoring function $H(\mathcal{M}, A + \alpha B)$, such that $|||P_1|| + ||\mathrm{H_2O}|| - MW_1| \leq \delta$, $|||P_2|| + ||\mathrm{H_2O}|| - MW_2| \leq \delta$. In the calculation, the molecular mass MW_1 or MW_2 is the neutral mass of the precursor without extra protons attached. We can obtain this value directly from the m/z and charge state z reported by the instruments. The scoring function $H(\mathcal{M}, A + \alpha B)$ measures how well the peptide pair (P_1, P_2) matches with the mixture spectrum $\mathcal{M}$, and A and B are the theoretical spectra predicted from their peptide sequence P_1 and P_2 respectively, and the coefficient α in the scoring function indicates the relative abundance of the co-sequenced precursors.

We use the *normalized dot product* of two real-value vectors to measure the spectral similarity in this research. When all the spectra are scaled to Norm 1, the *normalized dot product* will be simplified to be the calculation of the *cosine* value between two unit vectors. Such measurement scheme considers no special requirements regarding the real peak intensity values,

and the spectral similarity is measured based on the shape of the spectra being matched only. More specifically, we define the following way to convert each spectrum to a real-value vector. Assume that a spectrum $\mathcal{S}$ can be transformed to a real-value vector $V_{\mathcal{S}} = s_1, s_2, ...s_n$, in which each element s_i corresponds to the total intensity of peaks falling into the i^{th} mass bin. The value s_i is calculated as follows:

$$s_i = \sum_{(x_j,h_j)\in\mathcal{S}, x_j\in[(i-0.5)\delta,(i+0.5)\delta]} h_j \tag{4.1}$$

The bin size δ in the equation above is chosen according to the resolution of the instruments. Based on the analysis above, we provide in Algorithm 3 describing details of converting a MS/MS spectrum to its corresponding vector.

Algorithm 3 Converting an MS/MS spectrum to its vector representation

INPUT: Given a mixture spectrum $\mathcal{M} = \{(x_i, h_i)|i = 1, 2, ..., m\}$, a value MW representing the largest range of molecular mass values of the precursors which is determined by the m/z range of the mass spectrometer, and the bin size δ which is predefined by the resolution of the mass spectrometer.
OUTPUT: A real value vector $V_{\mathcal{M}} = s_1, s_2, ...s_n$, in which the subscript n symbolizes that the vector has n dimensions.

1: Calculate the dimension size n of vector according to $n = \lceil \frac{MW}{\delta} \rceil$.
2: Initializing each value $s_i = 0$ in $V_{\mathcal{M}}$
3: **for** Each peak (x_j, h_j) in $\mathcal{M}$, j from 1 to m **do**
4: $\quad index = \frac{x_j}{\delta}$
5: $\quad$ **if** $index - \lfloor index \rfloor < 0.5$ **then**
6: $\quad\quad i = \lfloor index \rfloor$
7: $\quad$ **else**
8: $\quad\quad i = \lceil index \rceil$
9: $\quad$ Add up the intensity value $s_i = s_i + hj$
10: Report the real value vector $V_{\mathcal{M}}$.

The bin size in Equation 4.1, and the error bound in the *Problem Formulation* are consistent with each other, both are predetermined by the resolution of the experimental configurations, therefore we use the same symbol δ to denote them. After the conversion, each vector is normalized to unit vector with each element in the vector divided by its Euclidean Norm, thus the scoring function can be rewritten in the following way:

$$H(\mathcal{M}, A + \alpha B) = \frac{V_{\mathcal{M}} \cdot (V_A + \alpha V_B)}{\|V_A + \alpha V_B\|} \tag{4.2}$$

In the equation above, the vector $V_{\mathcal{M}}$, V_A and V_B are all unit vectors. The $\|V_A + \alpha V_B\|$ in the denominator indicates the Euclidean Norm (or Euclidean Distance) of the new vector $V_A + \alpha V_B$, which is the linear combination of two separate unit vectors V_A and V_B.

4.3 Main Method

Effective computational approaches developed for the purpose of automated identification of mixture spectra will be beneficial for alleviating the bottleneck of low identification rate impeding the current computational analysis of mass spectrometry based proteomics data. Even though the proposed Mixture Spectra Database Search Problem is formulated in a very simple form, the direct implementation will suffer a major computational disadvantage when considering multiple precursors in one single spectrum. The number of the possible candidate peptide pairs that fall into the required mass error range will be very large. The computational burden for scoring each spectrum against all the possible peptide pairs will make it impractical for its application onto to large datasets. Moreover, the quadratic explosion in search space will also dramatically increase the chances of false-positive identifications. Under such circumstance, efficient filtration strategy is highly necessary for reducing the search space before scoring and ranking all the candidate peptide pairs.

4.3.1 Filtration Scheme

In the previous chapter, we formulated the problem of peptide *de novo* sequencing from mixture MS/MS spectra mathematically, and proposed a dynamic algorithm to report candidate pairs for each of the query spectrum. The algorithm has the ability to provide partially correct, yet useful peptide pairs for a given mixture spectrum. We will be utilizing these incomplete results to screen the peptide pairs acquired from the protein database.

Assume that for some mixture spectrum $\mathcal{M}$, the *de novo* algorithm will output the top-ranked peptide pairs in the following list:

$$L_m = \{(A_1, B_1), (A_2, B_2), ..., (A_m, B_m)\}$$

in which, the subscript m indicates the number of reported pairs, and it can be adjusted if required. Each element (A_i, B_i) in the list contains two individual peptides generated for the two different precursors in the query mixture spectrum respectively.

For each of the molecular mass values MW_1 and MW_2, we filtered the whole protein database to find all the theoretically digested peptide sequences that satisfy the required mass tolerance. We use the following list to include all the tentative peptide sequences after filtered by precursor mass value MW_1:

$$L_n^1 = (R_1, R_2, ..., R_n)$$

in which, each element R_i satisfy the requirement of mass tolerance δ such that $|||R_i|| + ||\mathrm{H_2O}|| - MW_1| \leq \delta$, and n is the number of the filtered sequences. Similarly we put all the possible peptide sequences for the second molecular mass value MW_2 in the following list:

$$L_k^2 = (Q_1, Q_2, ..., Q_k)$$

where for each element Q_j, the inequality $|||Q_j|| + ||\mathrm{H_2O}|| - MW_2| \leq \delta$ also holds.

Assuming that we intend to filter the peptide sequence list L_n^1: Firstly, for each element R_i, we compare it with all of its counterpart sequences in list L_m, that is to compare R_i with each A_j in the *de novo* candidate pairs. As for the filtration of the second sequence list L_k^2, similar comparison procedure also applies. Only difference is that we want to compare the sequence in L_k^2 with each sequence B_j in the *de novo* candidate list. Figure 4.1 gives an brief description of the filtration procedure.

Specifically, in the comparison of two sequences R_i and A_j, we use a special alignment algorithm which takes linear time to count the number of common amino acids $N_c^{(i,j)}$ between

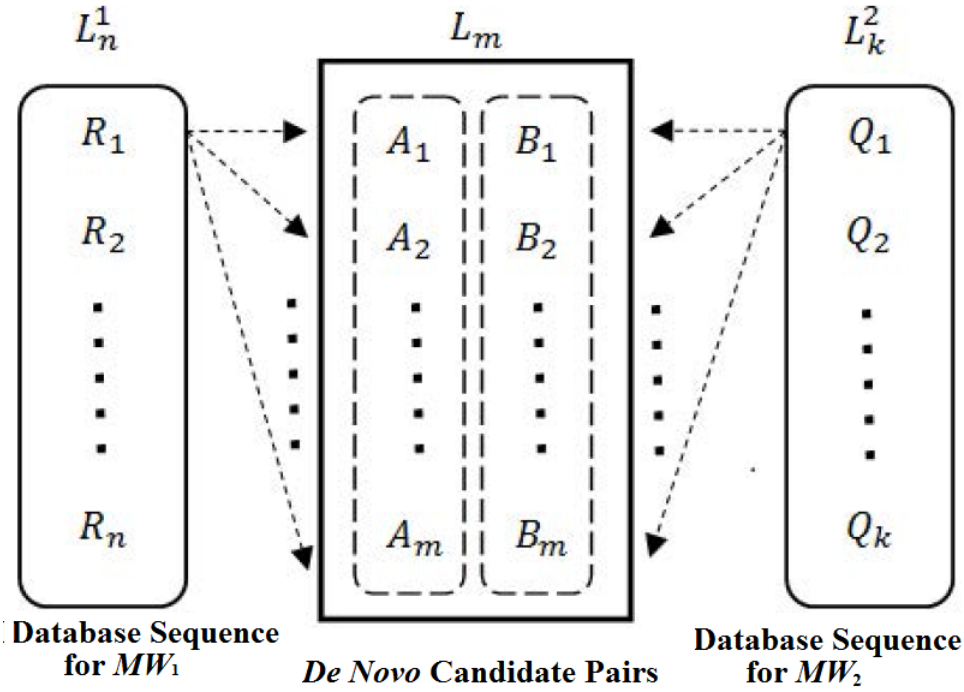

Figure 4.1: Exemplary illustration of the filtration procedure for database sequence lists L_n^1 and L_k^2. During such procedure, we compare each element sequence in L_n^1 or L_k^2 with their counterpart sequences in the *de novo* candidate list L_m.

them. An example to illustrate the comparison is shown in Figure 4.2, and the corresponding description for the alignment procedure is shown in Algorithm 4. In Algorithm 4, the procedure in Step 4 to Step 13 will at most be executed in $O(v + w)$ time where v and w are the lengths of the two input sequence respectively.

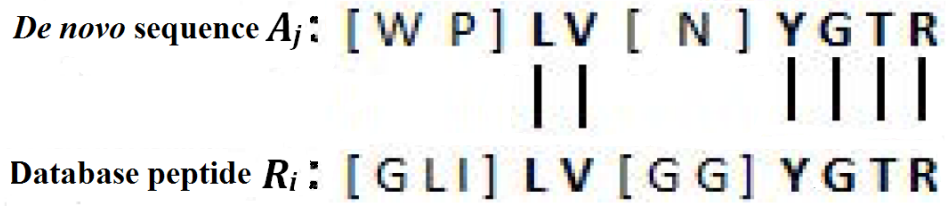

Figure 4.2: Comparing a *de novo* sequence with a database peptide. The alignment ensures that the mass of the aligned block(letters wrapped by brackets) is equal for both sequences. Although in this example the masses of [WP] and [GLI] are slightly different, we allow a tiny error tolerance δ exist, therefore we treat them as equal in comparison. The number of common amino acids here is $N_c = 6$.

Algorithm 4 Comparison between a *de novo* sequence with a database peptide sequence

INPUT: Given two sequences $A = a_1a_2...a_v$ and $R = r_1r_2...r_w$, and we have the tiny value δ which represent the predefined error tolerance.

OUTPUT: Find N_c which denotes the number of common amino acids between two input sequences. During the comparison, we assure that each matched or unmatched block between them are equal in weight.

1: Initialize $W_1 = 0$ and $W_2 = 0$ in which both are summations of the weight of amino acids.
2: Initialize $i = 1$ and $j = 1$ to be the indices in the computation
3: Assign an initial value for $N_c = 0$
4: **if** $|W_1 - W_2| \leq \delta$ **then**
5: **if** $a_i = r_j$ **then**
6: $N_c = N_c + 1$
7: $W_1 = W_1 + \|a_i\|$, $W_2 = W_2 + \|r_j\|$, $i = i + 1$, $j = j + 1$
8: **else**
9: $W_1 = W_1 + \|a_i\|$, $W_2 = W_2 + \|r_j\|$, $i = i + 1$, $j = j + 1$
10: **else if** $W_1 < W_2$ **then**
11: $W_1 = W_1 + \|a_i\|$, $i = i + 1$
12: **else**
13: $W_2 = W_2 + \|r_j\|$, $j = j + 1$
14: **if** $i \leq v$ and $j \leq w$ **then**
15: Repeat the procedure from Step 4
16: Output the value N_c which contains the number of common amino acids based on the special requirements stated above.

Secondly, we calculate an initial score for each R_i according to a *Triplet* $\mathcal{T}^i = (l^i_{max}, l^i_{sum}, m^i_{num})$ obtained during the comparison in the previous step. In the *Triplet*, the notation l^i_{max} represents the largest $N_c^{(i,j)}$ obtained when comparing the target peptide R_i with some sequence A_j from the *de novo* list L_m. And the notation l_{sum} is calculated as: $l^i_{sum} = \sum_{1 \leq j \leq n, N_c^{(i,j)} \geq 3} N_c^{(i,j)}$ which represents the summation over all the $N_c^{(i,j)}$ larger than or equal to 3. And the notation m^i_{num} denotes how many sequences in the *de novo* list L_m has common amino acids $N_c^{(i,j)} \geq 3$ when aligning to the current peptide R_i. The initial scoring function we choose in this filtration step is as follows:

$$S_{ini}(R^i) = \log(l^i_{sum}) * l^i_{max} \tag{4.3}$$

The filtration scoring function S_{ini} is chosen empirically and can be adjusted. Each of the three values contained in the *Triplet* indicates on some level how much the related peptide should be considered (or correct). Different scoring functions are evaluated based on the *Triplet*, however

we have observed similar performance in search space reduction after the filtration procedure. Another acceptable scoring function we have evaluated is $S_{ini}(R^i) = \log(\frac{l^i_{sum}}{m^i_{num}}) * l^i_{max}$.

Thirdly, we rank all the peptides in R_i based on the filtration score. The score calculated for each peptide indicates its likelihood of being a correct precursor, at least on some level. Similar operations can be carried out for the peptide list L^2_k. After scoring and ranking both L^1_n and L^2_k, we will select a portion of peptides from each list to form the candidate pairs which will be matched with the query mixture spectrum in the following step with a more rigorous scoring function. The algorithm for this part is described in Algorithm 5.

Algorithm 5 Filtration: Scoring and Ranking Database Peptides

INPUT: Given mixture spectrum $\mathcal{M}$, the list of *de novo* sequence pairs: $L_m = \{(A_1, B_1), (A_2, B_2), ..., (A_m, B_m)\}$, the database peptide list $L^1_n = (R_1, R_2, ..., R_n)$ for MW_1, and the database peptide list $L^2_k = (Q_1, Q_2, ..., Q_k)$ for MW_2.

OUTPUT: Both lists L^1_n and L^2_k sorted according to the scoring function S_{ini}

1: Initializing *Triplet* array $\mathcal{T}^1[1, ..., n]$ and $\mathcal{T}^2[1, ..., k]$.
2: Initializing the *Initial Score* array $S^1_{ini}[1, ..., n]$ and $S^2_{ini}[1, ..., k]$.
3: **for** i from 1 to n **do**
4: **for** j from 1 to m **do**
5: Calculate $N_c^{(i,j)}$ between database peptide R_i and *de novo* sequence A_j
6: **if** $N_c^{(i,j)} \geq l^i_{max}$ **then**
7: $l^i_{max} = N_c^{(i,j)}$
8: **if** $N_c^{(i,j)} \geq 3$ **then**
9: $m^i_{num} = m^i_{num} + 1$
10: $l^i_{sum} = l^i_{sum} + N_c^{(i,j)}$
11: Calculate the initial score $S^1_{ini}[i]$ for R_i based on Equation 3
12: Similar operations(line 3 to line 11)are carried out for L^2_k to obtain $\mathcal{T}^2$ and S^2_{ini}.
13: Sort both lists L^1_n and L^2_k in decreasing order according to the filtration score.

In line 5 of Algorithm 5, given the fact that both peptide R_i and sequence A_j have very limited length, the time complexity for finding the common amino acids between them can be regarded as a constant. Without losing generality, assume $n > k$, thus the overall complexity for Algorithm 5 is $O(mn + n \log n)$. m in the complexity denotes the number of candidate pairs reported by the *de novo* procedure, it can be easily adjusted according to real requirements of

balancing results accuracy and computational speed. In line 8 of Algorithm 5, we only count the case that the number of common amino acids are larger than 3. because larger number indicates more confidently that those common letters are true matching evidence rather than random hits between the aligned sequences.

After the database peptides are sorted, we calculate the ratio between the first-ranked and second-ranked sequences in each database peptide lists. We use ratio $r = \frac{R_{1st}}{R_{2ed}}$ to determine how many peptides in each list should go through further examination. A relatively larger value of r strongly suggests that the top-ranked peptide has a great chance of being one of the co-fragmented precursors. If $r - 1 \geq \beta$ in which β is a threshold value satisfies $\beta \geq 0$, we only have to take out the very few top-ranked peptides. The threshold we use in this research is $\beta = 0.3$, in case that $r - 1 \geq \beta$ we only select the top $\lfloor \log n \rfloor$ peptides out of the list, otherwise if $r - 1 < \beta$ we will take out all the peptides in the list which has $m^i_{num} > 0$. The threshold β is also an empirical value in the research. Base on our preliminary experiment on a dataset of limited size, we found that $\beta = 0.3$ is a threshold value large enough to distinguish the first-ranked peptide from its followers. Peptides with $m^i_{num} > 0$ means that there is at least one *de novo* sequence that has 3 or more common amino acids with the current database peptide. Thus in case that we can't rely on the ratio r to reduce the list, we will consider all the peptides with such *de novo* matching evidence. This also helps to cut down the number of peptides to be considered in the next step. After the operations above, we will pair up the peptides selected from different lists to form peptide pairs and score each peptide pair against the query mixture spectrum based on Equation 4.2 to report the best matched pair.

4.3.2 Estimation of Mixture Coefficient

In the query mixture spectrum $\mathcal{M} = A + \alpha B$, we assume that α is unknown before the identification. It is necessary to give a reasonable estimation of the coefficient α prior to scoring the

query mixture spectrum against the target candidate pair, because a biased mixture coefficient will compromise the accuracy of calculating the *normalized dot product* between $\mathcal{M}$ and the correct peptides. We use a similar method as [114] to estimate the mixture coefficient. Assume that α' denotes the estimated value of the mixture coefficient, we obtain the optimal value of α' such that the *cosine* similarity between $V_{\mathcal{M}}$ and $V_A + \alpha' V_B$ is maximized. Because vectors $V_{\mathcal{M}}$, V_A and V_B are all normalized to unit vectors, then $V_{\mathcal{M}}^2 = V_A^2 = V_B^2 = 1$. We rewrote the mixture spectra scoring function in Formula 2 as the following function with respect variable α:

$$f(\alpha) = \frac{V_{\mathcal{M}} \cdot (V_A + \alpha V_B)}{\sqrt{1 + 2\alpha V_A \cdot V_B + \alpha^2}}$$

in which $f(\alpha) = \cos\beta$, and β is the angle between the vectors $V_{\mathcal{M}}$ and $V_A + \alpha V_B$.

The function $f(\alpha)$ will have a maximum value at some value α'. In order to achieve this, the first derivative of $f(\alpha)$ with respect to α will be zero at this specific α', meanwhile the second derivative of $f(\alpha)$ at the corresponding α' is negative. To simplify the following derivations, we denote $V_{\mathcal{M}} \cdot V_A = x$, $V_{\mathcal{M}} \cdot V_B = y$ and $V_A \cdot V_B = z$. The first derivative of function $f(\alpha)$ is calculated as:

$$f'(\alpha) = \frac{y\sqrt{1 + 2\alpha z + \alpha^2} - (x + \alpha y)(\alpha + z)(1 + 2\alpha z + \alpha^2)^{-\frac{1}{2}}}{1 + 2\alpha z + \alpha^2}$$

In which, the denominator will always be greater than zero, therefore to make $f'(\alpha) = 0$ is equivalent to let the numerator be zero, then we will have:

$$\begin{aligned} y\sqrt{1 + 2\alpha' z + \alpha'^2} - (x + \alpha' y)(\alpha' + z)(1 + 2\alpha' z + \alpha'^2)^{-\frac{1}{2}} &= 0 \\ y(1 + 2\alpha' z + \alpha'^2) - (x + \alpha' y)(\alpha' + z) &= 0 \\ y + 2\alpha' yz + \alpha'^2 y - \alpha' x - xz - \alpha'^2 y - \alpha' yz &= 0 \\ y - xz + \alpha' yz - \alpha' x &= 0 \end{aligned}$$

From the induction above, we will obtain the following formula to calculate the estimation of mixture coefficient:

$$\alpha' = \frac{y - xz}{x - yz} = \frac{V_{\mathcal{M}} \cdot V_B - (V_{\mathcal{M}} \cdot V_A)(V_A \cdot V_B)}{V_{\mathcal{M}} \cdot V_A - (V_{\mathcal{M}} \cdot V_B)(V_A \cdot V_B)} \tag{4.4}$$

In our model, the mixture spectrum is modelled as the linear combination of two individual spectra in the form $\mathcal{M} = A + \alpha B$, therefore the the mixture coefficient that represents the relative abundance level of two co-sequenced peptides should be at least positive. However, from Equation 4.4, we cannot guarantee the value α' calculated here is always positive, theoretically it could have three different cases. Figure 4.3 illustrated the three different cases for α' followed by the corresponding explanation for each of them.

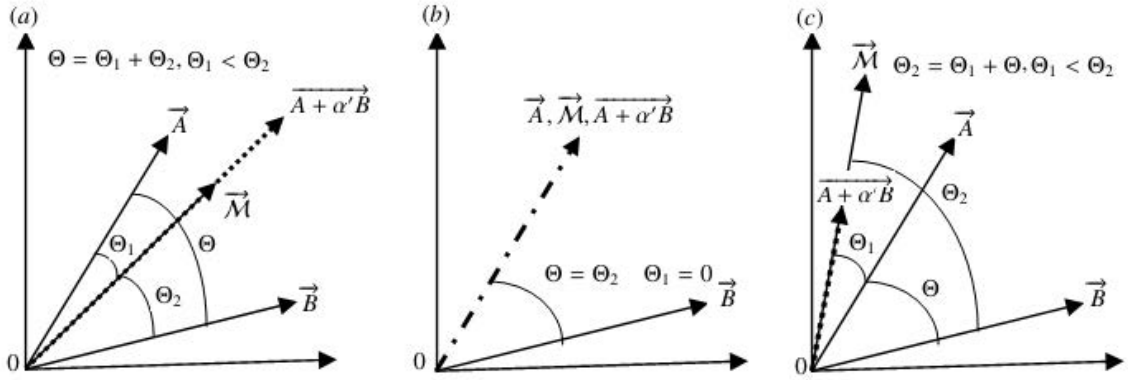

Figure 4.3: Exemplary figures describing three different cases that the value α' can be. In (a), it satisfies $\alpha' \in (0, 1)$. In (b), $\alpha' = 0$. In (c), it satisfies $\alpha' < 0$.

Without losing generality, we assume that $x > y$ which means $\overrightarrow{\mathcal{M}}$ is more similar to $\overrightarrow{A}$ than it is to $\overrightarrow{B}$, and it also indicates that the angle Θ_1 between $\overrightarrow{\mathcal{M}}$ and $\overrightarrow{A}$ is smaller than the angle Θ_2 between $\overrightarrow{\mathcal{M}}$ and $\overrightarrow{B}$, $\Theta_1 < \Theta_2$. As shown in Figure 4.3, the angle between $\overrightarrow{A}$ and $\overrightarrow{B}$ is denoted as Θ.

In (a) of Figure 4.3, the target vector $\overrightarrow{\mathcal{M}}$ is somewhere between $\overrightarrow{A}$ and $\overrightarrow{B}$, thus we can find a proper value α' such that the newly constructed vector $\overrightarrow{A + \alpha' B}$ will parallel with the target vector $\overrightarrow{M}$, and only at this specific value the *normalized dot product* between $\overrightarrow{\mathcal{M}}$ and $\overrightarrow{A + \alpha' B}$ is maximized. It is expected that $\alpha' \in (0, 1)$ in this case. This example describes a scenario that the target spectrum $\mathcal{M}$ is highly likely to be a mixture of spectra A and B.

In (b) of Figure 4.3, the target vector $\overrightarrow{\mathcal{M}}$ overlaps with $\overrightarrow{A}$ which describe a theoretical

case that vectors $\overrightarrow{\mathcal{M}}$ and $\overrightarrow{A}$ are indeed identical, and the value for α' that will maximize the *normalized dot product* between between $\overrightarrow{\mathcal{M}}$ and $\overrightarrow{A + \alpha' B}$ is $\alpha' = 0$. This example tells that the target spectrum $\mathcal{M}$ is generated from a single precursor instead of any mixture of A and B.

In (c) of Figure 4.3, the target vector $\overrightarrow{\mathcal{M}}$ is outside the sector area enclosed by $\overrightarrow{A}$ and $\overrightarrow{B}$. Theoretically, there also exists a value of α' such that the newly generated vector $\overrightarrow{A + \alpha' B}$ parallels with $\overrightarrow{\mathcal{M}}$ which actually maximizes the required *normalized dot product*. Noticeably, we can see that this α' will be a negative value. This example indicates that the target spectrum $\mathcal{M}$ is not any mixture of A and B, and it is more likely to be generated by A.

In practice, when the value for α' calculated by Equation 4.4 is $\alpha' < 0$, or any value that has the approximation $\alpha' \approx 0$, then we consider the query spectrum $\mathcal{M}$ to be a single precursor spectrum instead of a mixture spectrum. If α' is in the range $(0, 1)$, then we simply treat $\mathcal{M}$ as a mixture/linear combination of A and B.

Besides calculating the estimation of mixture coefficient, to ensure that at this specific point α' computed by Equation 4.4, the original function $f(\alpha)$ obtains the maximum value, we also need to guarantee that the second derivative of $f(\alpha)$ is negative at α'. This can be obtained by the following deductions: We know that $f(\alpha) = \cos\beta$ will always be positive, thus $\beta \in (0, \frac{\pi}{2})$, the second derivative of this function is $-\cos\beta$. In the domain $\beta \in (0, \frac{\pi}{2})$, it will ways be negative value for $-\cos\beta$. Thus, we can conclude that the value α' calculated by Equation 4.4 will always make the original function $f(\alpha)$ achieve the maximum value.

4.3.3 Algorithm for Scoring Peptide Pairs

Given an query mixture spectrum $\mathcal{M}$, our proposed method will firstly conduct a preliminary *de novo* procedure to report a list of sequence pairs, and then those *de novo* results possessing only

limited accuracy will be used to filtrate the database peptides acquired from a protein sequence database. After filtration of the database peptides, it is expected that the number of potential peptide sequences that require a more rigorous examination will be reduced. For each peptide sequence in the shortened lists, we use a similar method as [11] to predict its corresponding theoretical spectrum. All the fragment types considered in the theoretical spectrum are denoted

Algorithm 6 Scoring Peptide Pairs against Query Mixture Spectrum

INPUT: The query mixture spectrum $\mathcal{M}$ and shortened database peptide lists after filtration: $L_x^1 = (R_1, R_2, ..., R_x)$ for precursor molecular value MW_1, and $L_y^i = (Q_1, Q_2, ..., Q_y)$ for precursor molecular value MW_2.

OUTPUT: The best matched peptide pair (R_i, Q_j) and its matching score.

1: Initializing an variable $Score_{\max} = 0$, and two indices $x_{\max}$ and $y_{\max}$
2: Convert $\mathcal{M}$ to vector and normalize to unit vector $V_{\mathcal{M}}$.
3: **for** i from 1 to x **do**
4: Predict theoretical spectrum S_{R_i} from sequence R_i
5: Convert S_{R_i} and normalize to unit vector V_{R_i}
6: **for** j from 1 to y **do**
7: Predict theoretical Spectrum S_{Q_j} from sequence Q_j
8: Convert S_{Q_j} and normalize to unit vectors V_{Q_j}
9: Estimate coefficient α' based on Equation 4.4
10: Calculate the *cosine* value using $\cos = \frac{V_{\mathcal{M}} \cdot (V_{R_i} + \alpha' V_{Q_j})}{\|V_{R_i} + \alpha' V_{Q_j}\|}$
11: **if** $\cos > Score_{\max}$ **then**
12: $Score_{\max} = \cos$
13: $x_{\max} = i$ and $y_{\max} = j$
14: Output peptide pair $(R_{x_{\max}}, Q_{y_{\max}})$ with its matching score $Score_{\max}$

in $\Pi = \{y, b, a, c, x, z, y^*, y^o, b^*, b^o\}$. The theoretical spectrum is furthermore converted to a spectrum vector using Equation 4.1 and subsequently normalized to a unit vector. Sequences from two individual lists are paired up to constitute the tentative peptide pairs. Then for each of the tentative peptide pairs, we will score it against the query mixture spectrum based on the scoring function in Equation 4.2, and the best matched pair will be outputted in the final step. Algorithm 6 describes the general outline of seeking the best matched pair among all the tentative pairs.

The time consumed in Line 2, 5, 8 for converting a vector to its corresponding unit vector depends on the number of bins considered in the spectra conversion, therefore the complexity

is $O(\frac{MW}{\delta})$. The overhead in Line 10 for calculating the *normalized dot product* between $V_{\mathcal{M}}$ and the linear combined vector $V_{R_i} + \alpha' V_{Q_j}$ also relies on the number of bins(or dimensions) in the vector, thus the complexity for this part is the same as above. Meanwhile, we have nested loops iterative with i and j, therefore the integrated complexity for Algorithm 6 is $O(x \times y \times \frac{MW}{\delta})$, in which x and y are denoted in Algorithm 6 representing the size of the shortened database peptide lists respectively.

4.4 Experiment Result and Discussion

4.4.1 Experiment Summary

Our previous attempt of *de novo* sequencing from mixture spectra laid the foundation for the research in this manuscript. A preliminary *de novo* sequencing procedure with only partial correctness is integrated in the filtration strategy to reduce the examination space. Prior to scoring the query mixture spectra against the enormous amount of candidate pairs by the stringent scoring function in Formula 4.2, the *de novo* assisted filtration strategy will initially rank all the peptides that are obtained directly from the theoretical digestion of the protein sequence database by a preliminary scoring model and only those higher-ranked peptide sequences are selected and paired up to form candidate pairs that can go through the subsequent stringent scoring procedure.

To verify the efficiency of the proposed method, we use the published software package MSPLIT [114] on a tryptic yeast dataset released on PRIDE [123] repository to obtain some real mixture spectra. The MSPLIT software is run in the target/decoy strategy on the yeast dataset collected from an Ion Trap mass spectrometer. In our experiment, we further filtered those reported mixture spectra according to our current requirements. We only select those spectra in which both precursor peptides in the mixture spectra have charge 2, and contain no

Post-translational Modifications. Furthermore, we use the renowned database search software PEAKS DB [10] to re-confirm those mixture spectra reported by MSPLIT. Each individual peptide is searched against the protein database with the other peptide sequence removed from the database to eliminate the potential interference between two peptides. Only those mixture spectra in which both peptides can be confirmed by PEAKS DB are kept. In total, we obtained 7 distinct mixture spectra. We implemented a software prototype based on our proposed method

Table 4.1: Preliminary experiment results on a dataset containing 7 mixture spectra. The columns N_b and N_a represent the number of database peptides before and after the filtration procedure for each mass value respectively. The column f (in millesimal) shows the ratio between the number of candidate sequence pairs after filtration and the number of all possible peptide pairs acquired directly from the protein database. The column α' is the estimated mixture coefficient. The column $\cos\Theta$ is the score (*normalized dot product*) calculated based on Equation 4.2

m/z	Peptides	N_b	N_a	f	α'	$\cos\Theta$
553.321	GLILVGGYGTR	1853	7	0.5831‰	0.827	0.327
553.7795	GPPGVFEFEK	1924	297			
419.724	ASIASSFR	1823	7	0.0204‰	1.101	0.444
420.2585	IAGLNPVR	1319	7			
506.78	FHLGNLGVR	1709	7	0.0156‰	0.917	0.365
507.2820	KFPVFYGR	1837	7			
600.2756	GYSTGYTGHTR	976	6	0.0292‰	1.149	0.321
600.340	DAGTIAGLNVLR	1476	7			
462.706	DNEIDYR	1239	474	267.23‰	1.347	0.343
463.3077	IVAALPTIK	743	519			
521.253	YSDFEKPR	1434	7	0.0182‰	0.996	0.302
521.7932	GAIAAAHYIR	1880	7			
675.364	GKPFFQELDIR	1560	360	1.0147‰	1.847	0.388
675.8437	ANLGFFQSVDPR	1592	7			

and searched those mixture spectra against a yeast protein sequence database. The theoretically enzymatic peptides acquired from the protein sequence database contain both fully-tryptic and semi-tryptic peptide sequences, and also contain the peptide sequences with one missing cleavage. The mass error tolerance considered throughout the experiment is $\pm 0.1Da$. Our software can successfully identify all the 14 different peptides contained in those mixture spectra above. The experimental results are listed in Table 4.1.

From Table 4.1, we can clearly see that the proposed filtration strategy can effectively reduced the number of candidate peptide pairs to be examined. For most of the entries, the reduction ratio f is less than 1‰ or we say less than one thousandth. Even with the worst case in entry 5, the proposed method can still exclude more than two thirds of the sequence pairs acquired from the protein sequence database after filtration. We noticed that in this entry, the filtration doesn't perform as good as others. It is mainly because that after the initial ranking of both tentative peptide lists, the first ranked peptide has a similar initial score value as the second ranked peptide in either of the lists, thus they are not easy to be distinguished. In this case, we use a relatively larger set in order to include the correct peptide for further calculation. To be more specific, all the peptide candidates that have $m^i_{num} > 3$ are included in the set which will be then paired up and matched with the query mixture spectrum. Another point worth noticing is that in the real mixture spectra, we don't know which precursor have the higher abundance, thus the estimated value for α' can be either larger than 1 or smaller than 1. The correctness of the identification results demonstrated the effectiveness of the proposed method for matching a mixture spectrum with a pair of database peptides by using the filtration strategy. In our future work, we will evaluate the performance of our proposed method on datasets of different size and different acquisition methods, also we will develop a method for the validation of the identification results.

4.4.2 Annotation of Identifications

To provide a better view of the mixture spectra identification results, our software prototype enables generating the graphical annotations for the identified mixture MS/MS spectra. The simple interface enables users to zoom into (or zoom out) a specific locations on both the horizontal direction of m/z values. Dynamically, when pointing to a specific peak in the plot, the corresponding m/z and intensity values of that peak will be simultaneous displayed. In each annotated spectrum, the matched ion fragments for the two identified peptides are both labelled.

The software prototype has the ability to report and annotate more than one matched peptide pairs, and by default the top ranked peptide pair for each query mixture spectrum is outputted as the identification results. An example of the spectrum annotation is shown in Figure 4.4. The annotated spectra will provide an extra dimension of information on the identification results as well as the matched ion fragments in the spectra. From Figure 4.4,we can see there are quite a few labels that contains multiple annotations which indicates that in the mixture MS/MS spectra the occurrence of overlapping fragments can be frequent. Appendix A provides the annotations of the results for all the other identified mixture spectra in the experiment.

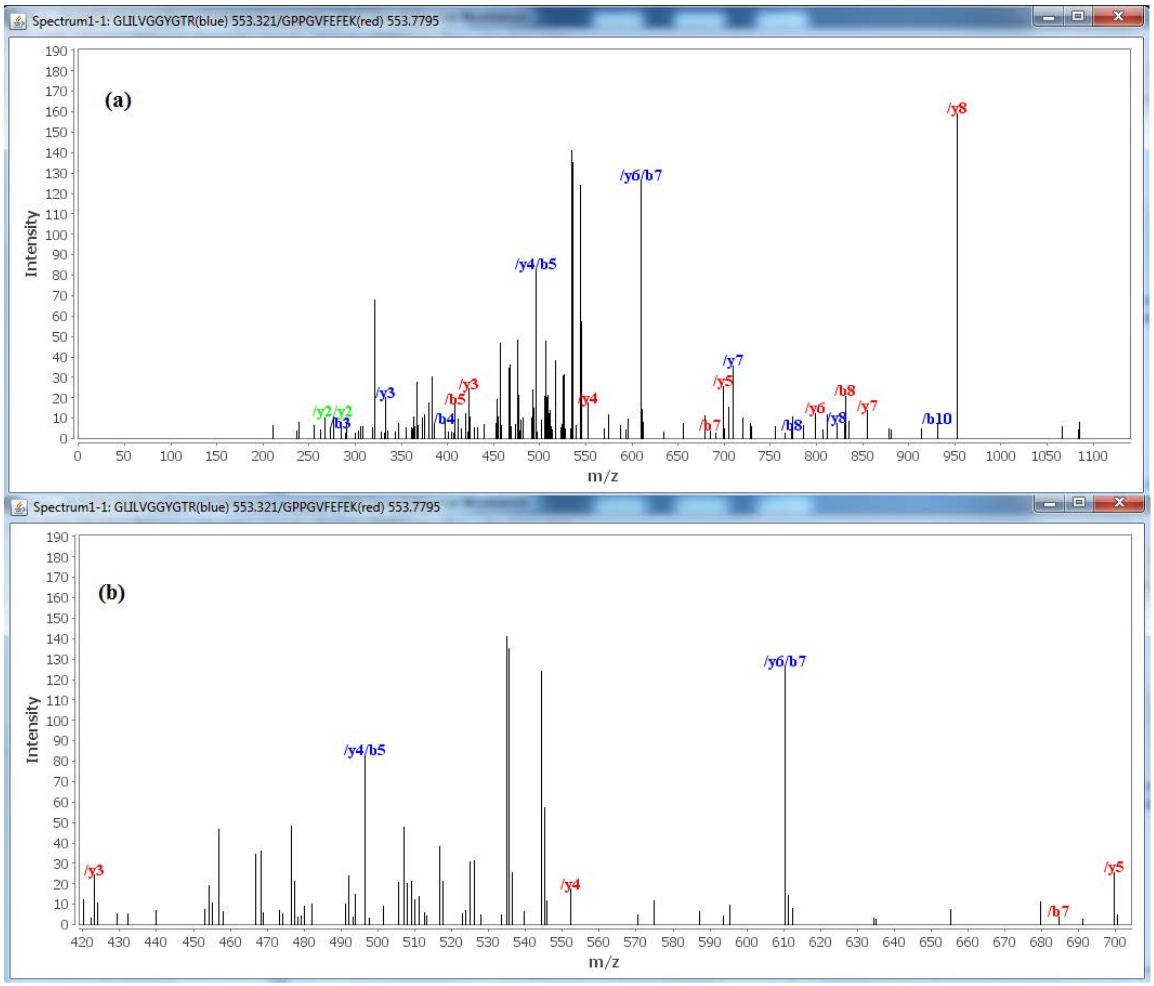

Figure 4.4: Example of the spectrum annotations. The reported identifications for the query mixture spectrum are **GLILVGGYGTR** and **GPPGVFEFEK** shown in **(a)**. In the annotated spectrum, the blue signs are the matched ion fragments from the first peptide, and the red ones are the matched ions of the second peptide, in the meanwhile the green ones in the spectrum are the reported overlapping peaks. When zooming in around the value $m/z = 550$, we can see more details in the specific location in **(b)**.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Steady progresses have been achieved in the applications of mass spectrometry in proteomics studies in recent years. However, one frustrating fact that impedes the efficiency of computational approaches is the low identification rate of the acquired mass spectral data. The existence of mixture spectra is an influential factor contributing to this situation, which also significantly complicates the analysis of the mass spectrometry data, yet their effects on peptide and protein identification in complex mixtures remains unveiled. There is currently an increasing necessity of developing effective approaches for the characterization of mixture spectra. The successful identification of mixture spectra will directly contributes to improving the overall throughput of proteomics experiments, especially when analyzing protein sample with increasing complexity. Also, along with the arising of new acquisition protocols, the occurrence of mixture spectra is frequent, even intentional which makes such algorithms and software tools for analyzing mixture spectra more desired. Thus, the almost ubiquitous assumption that every MS/MS spectrum comes from one precursor becomes unsuitable to identify spectra from co-eluting peptides. In this manuscript, we conducted research regarding the identification of mixture spectra on two different yet highly correlated topics: the *de novo* sequencing of mixture spectra and mixture

spectra database search.

To the best of our knowledge, our research on the mixture MS/MS spectra *de novo* sequencing is the first effort of its kind. We proposed a sophisticated dynamic programming algorithm for the purpose of reporting a pair of peptide sequences for a query mixture spectrum. Unlike the previous attempts of seeking multiple peptides from a single mixture spectrum in an iterative manner, our method modelled the mixture spectrum as the linear combination of ion fragments from both co-sequences precursors and proceeded in an carefully designated pathway to reconstruct both peptides concurrently. The effectiveness of the proposed algorithm is benchmarked on both simulated and real mixture spectra datasets. Additionally, in the research we provide two scenarios for the possible applications of the proposed method. One scenario is based on the fact that we formulate the concurrent fragmented precursors together, and as demonstrated in the simulated mixture spectra, this provides the proposed method the ability to report better results in the case that the co-sequenced precursors have comparable abundances. In this scenario, if we can find a mixture spectrum in the dataset in which the co-sequenced precursors have relatively approximate intensities, our approach can be applied in this situation to report more useful results. Another scenario relies on the traditional idea of integrating *de novo* sequencing results with database search method to accurately and sensitively identify MS/MS spectra. As shown in the simulated mixture spectra, our proposed method has a good chance to report useful though partially correct peptide sequences, and those partially correct results are highly useful in reducing the examination space during the database search.

The second application scenario above laid the foundation for the subsequent research on mixture spectra database search. In our most recent research, we combined the *de novo* results with database search to identify the acquired mixture spectra. The *de novo* results here provide a way to score and rank those peptide candidates from the theoretical digestion of protein database. After ranking against the preliminary *de novo* sequences, only those database peptides with larger initial filtration score are filtered out and go through a more stringent

scoring procedure in which calculating the similarity between the experimental spectrum and the theoretical spectrum is transformed to computing the *normalized dot product* of two real-value vectors. Such unique strategy of filtering the candidate peptide pairs provides us a major improvement in efficiency by which several magnitudes of reduction in search space is obtained. Besides the filtration strategy, we use the mixture coefficient α to represent the relative abundance of the precursors, and we also introduced a method to estimate the relative abundance level of the co-sequenced precursors. Previous reports indicate that the abundances of fragments are indeed highly related to the abundance of the corresponding precursors, thus it is necessary to give an appropriate estimation of the mixture coefficient prior to the rigorous scoring procedure and an improper mixture coefficient will surely detriment the accuracy of identification. This special *de novo* assisted database search method shows great efficiency on our preliminary experiment result, in which the correct peptide pairs can be identified for each query mixture spectra by only searching through a tiny fraction of all the peptides candidate pairs acquired directly from the protein sequence database.

5.2 Future Work

Even with all the advances accomplished in the thesis, the research on characterizing mixture spectra remains unsatisfactory. Just like what happened in dealing with simple peptide spectra, the proteomics community will have to make continuous efforts on the problem of understanding mixture spectra before we are able to unveil the underlying myth. For the specific problem of mixture spectra identification, we will concentrate on several research topics described in the following part.

First, we will continue and improve our current work on the *de novo* sequencing of mixture spectra and its related applications. More efficient algorithms will be needed to improve the identification accuracy, as well as to reduce the time and space complexity. The major problem

that affects the practicability of *de novo* approach is the low identification accuracy. In order to alleviate such bottleneck, another promising research work that we want to conduct is to find the sequence tags instead of whole sequence by *de novo* approach. The *de novo* sequence tags with high confidence level will be more useful in helping the database search to reduce the examination space. We can also combine such algorithmic solutions with database search techniques to obtain better identification results of mixture spectra.

Second, we will conduct research on the accurate determination of monoisotopic mass values of the co-sequenced precursors. The mass spectral peak representing the monoisotopic mass is sometimes not the most abundant isotopic peak in a spectrum despite it contains the most abundant isotope for each atom. This is because as the number of atoms in a molecule increases, the probability that entire molecule contains at least one heavy isotope atom also increases. Mixture spectra are generated from the co-fragmentation of multiple peptides. One complication here is to separate the monoisotopic precursors of these peptides from each other. Conventionally, the identification of MS/MS spectra is closely related to the monoisotopic masses of the precursors, it is reasonable to conclude that the more accurate the precursor masses are, the more easily and accurately the spectra can be identified. The monoisotopic peak is sometimes not observable for two primary reasons. One case is that the monoisotopic peak may not be resolved from other isotopic peaks or the monoisotopic peak is below the noise level. The other case is that the monoisotopic peak is not included in the selection window, this occurs both in the *Data Dependent Acquisition* and *Data Independent Acquisition* modes. All these special circumstances indicate that the precursor mass value we acquired from the mass spectrometer may not be the monoisotopic value. Correctly determining the monoisotopic peaks of the co-fragmented precursors in the isotopic clusters will help to increase the identification accuracy of the collected spectra.

Third, we will try to combine our previous research on the *de novo* sequencing research with the spectral library search to identify mixture spectra. The advances in high-throughput

MS/MS have promoted the accelerated growth of publicly available libraries of single peptide spectra, comprising of large volume collection of peptide MS/MS spectra with validated identification results. And the availability of these large spectral archives has ignited the research on developing spectra identification approaches based on spectral matching and alignment algorithms. Prior to the spectral library searching process, we can conduct a preliminary *de novo* sequencing procedure to retrieve some useful information regarding the target peptides. On one hand those information, either sequence tags or partially correct sequences will be beneficial to screen the spectral candidates from the spectral library, on the other hand, if the target spectrum and peptide is not observed in the library, the selected sequences identified from mass spectra with good *de novo* reconstructions will be considered as novel peptides.

Last but not least, solving the problem of peptide identification from mixture spectra represents an important step toward addressing related and emerging problems in proteomics, among which a specific topic is to identify the disulphide linked peptides. Similar to the co-fragmentation of multiple precursors that generates mixture spectra, the covalently linked peptides, such as disulphide bonded peptides also give rise to the generation of non-canonical MS/MS spectra. The formation of disulphide bonds is critical for stabilizing protein structures and maintaining protein functions, therefore knowledge of the disulphide linkage will provide a deeper understanding of the tertiary structure and biological function of proteins. The peptides linked by disulphide bond are co-sequenced together in the LC-MS/MS, which will produce the spectra containing fragments from both peptides. A straightforward yet reasonable model for formulating the problem is to treat the disulphide linked peptides as a tree. We consider the bonded location between two peptides as the root of the tree, and starting from the root, the tree has four different branches which corresponds to the prefixes and suffixes of two different peptides. By isolating the correct peaks in the spectrum that match with the theoretical mass values of ion fragments, we may be able to reconstruct the covalently linked peptide sequences and locate the disulphide bonds in a *de novo* sequencing manner. We can also design efficient

algorithmic solutions to match the experimental derived spectra with peptide sequences from database, and to subsequently determine the number and locations of the disulphide bonds, this is similar to the database search of mixture spectra.

Bibliography

[1] Blackstock, W.P. and Malcolm, P.W.: Proteomics: Quantitative and Physical Mapping of Cellular Proteins. Trends in Biotechnology, 17(3), 121-127, 1999

[2] Petricoin, E.F., Zoon, K.C., Kohn, E.C., Barrett, J.C. and Liotta, L.A.: Clinical Proteomics: Translating Benchside Promise into Bedside Reality. Nature Reviews Drug Discovery, 1(9), 683-695, 2002

[3] Aebersold, M. and Mann, M.: Mass Spectrometry-based Proteomics. Nature, 422(6928), 198-207, 2003

[4] Domon, B. and Aebersold, R.: Mass Spectrometry and Protein Analysis. Science, 312(5771), 212-217, 2006

[5] Karas, M., Bachmann, D., Bahr, U.E. and HillenKamp, F.: Matrix-assisted Ultraviolet Laser Desorption of Non-volatile Compounds. International Jounal of Mass Spectrometry and Ion Processes, 78, 53-68, 1987

[6] Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse C.M.: Electrospray Ionization for Mass Spectrometry of Large Biomolecules. Science, 246(4962), 64-71, 1989

[7] Peng, J., Elias, C.C., Thoreen, L.J., Licklider, L.J. and Gyqi, S.P.: Evaluation of Multidimentional Chromatography Coupled With Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-scale Protein Analysis: The Yeast Proteome. Jounal of Proteome Research, 2(1), 43-50, 2003

[8] Papping, D.J.C., Hojrup, P. and Bleasby, A.J.: Rapid Identification of Proteins by Peptide-Mass Fiingerprinting. Current Biology, 3(6), 327-332, 1999

[9] Cottrell, J. S., and London, U.: Probability-based Pprotein Identification by Searching Sequence Databases using Mass Spectrometry Data. Electrophoresis, 20(18), 3551-3567, 1999

[10] Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. and Ma, B.: PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. Molecular and Cellular Proteomics, 11(4), M111-010587, 2012

[11] Eng, J.K., Mccormack, A.L. and Yates, J.R.: An Approach to Correlate Tandem Mass-spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. Journal of the American Society for Mass Spectrometry, 5(11), 976-989, 1994

[12] Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W. and Bryant, S.H.: Open Mass Spectrometry Search Algorithm. Journal of Proteome Research, 3(5), 985-964, 2004

[13] Craig, R. and Beavis, R.C.: TANDEM: Matching Proteins with Tandem Mass Spectra. Bioinformatics, 20(9), 1466-1467, 2004

[14] Colinge, J., Masselot, A., Giron, M., Dessingy, T. and Magnin, J.: OLAV: Towards High-throughput Tandem Mass Spectrometry Data Identification. Proteomics, 3(8), 1454-1463, 2003

[15] Bartels, C.: Fast Algorithm for Peptide Sequencing by Mass Spectroscopy. Biomedical and Environmental Mass Spectrometry, 19(6), 263-368, 1990

[16] Taylor, J.A. and Johnson, R.S.: Implementation and Uses of Automated De Novo Peptide Sequencing by Tandem Mass Spectrometry. Analytical Chemistry, 73(11), 2594-2604, 2001

[17] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A and Lajoie, G.: PEAKS: Powerful Software for Peptide De Novo Sequencing by Tandem Mass Spectrometry. Rapid Communication in Mass Spectrometry, 17(20), 2337-2342, 2003

[18] Frank, A. and Pevzner, P.: PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modelling, Analytical Chemistry, 77(4), 964-973, 2005

[19] Chi, H., Sun, R.X., Yang, B., Song, C.Q., Wang, L.H., Liu, C., Fu, Y. Yuan, Z.F., Wang, H.P, He, S.M. and Dong, M.Q.: pNovo: De Novo Peptide Sequencing and Identification Using HCD Spectra. Journal of Proteome Research, 9(5), 2713-2724, 2010

[20] Bern, M. and Goldberg, D.: De Novo Analysis of Peptide Tandem Mass Spectra by Spectral Graph Partitioning. Journal of Computational Biology, 13(2), 364-378, 2006

[21] Dancik, V., Theresa, A.A., Karl, R.C., James, E.V. and Pevzner, P.A.: De Novo Peptide Sequencing via Tandem Mass Spectrometry. Journal of Computational Biology, 6(3-4), 327-342, 1999

[22] Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W. and Buhmann, J.M.: NovoHMM: A Hidden Markov Model for De Novo Peptide Sequencing. Analytical Chemistry, 77(22), 7265-7273, 2005

[23] Mo, L., Debojyoti, D., Wan, Y. and Chen, T.: MSNovo: A Dynamic Programming Algorithm for De Novo Peptide Sequencing via Tandem Mass Spectrometry. Analytical Chemistry, 79(13), 4870-4878, 2007

[24] Grossmann, J., Franz, Roos, F.F., Cieliebak, M., Liptak, Z., Mathis, L.K., Muller, M., Baginsky, S.: AUDENS: A Tool for Automated Peptide De Novo Sequencing. Journal of Proteome Research, 4(5), 1768-1774, 2005

[25] Demine, R. and Walden, P.: Sequit: Software for De Novo Peptide Sequencing by Matrix-Assisted Laser Desorption/Ionization Post-Source Decay Mass Spectrometry. Rapid Communications in Mass Spectrometry, 18(8), 907-913, 2004

[26] Day, R.M., Borziak, A. and Gorin, A.: PPM-chain: De nNovo Peptide Identification Program Comparable in Performance to Sequest. IEEE Computational Systems Bioinformatics Conference, Proceedings, 505-508, 2004

[27] Chen, T., Kao, M.Y., Tepel, M., Rush, J. and Church, G.M.: A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry. Journal of Computational Biology, 8(2), 325-337, 2001

[28] Okumura, N., Besada, V., Betancourt, L., Padron, G., Shimonishi, Y. and Takao, T.: Automated Interpretation of Low-energy Collision-induced Dissociation Spectra by SeqMS, a Software Aid for De Novo Sequencing by Tandem Mass Spectrometry. Electrophoresis, 21, 1694-1699, 2000

[29] DiMaggio, P.A. and Christodoulos, A.F.: De Novo Peptide Identification via Tandem Mass Spectrometry and Integer Linear Optimization. Analytical Chemistry. 79(4), 1433-1446, 2007

[30] Zheng, Z.: De Novo Peptide Sequencing based on a Divide-and-Conquer Algorithm and Peptide Tandem Spectrum Simulation. Analytical Chemistry, 76(21), 6374-6383, 2004

[31] Olson, M.T., Epstein, J.A. and Yergey, A.L.: De Novo Peptide Sequencing using Exhaustive Enumeration of Peptide Composition. Journal of the American Society of Mass Spectrometry, 17(8), 1041-1049, 2006

[32] Yergey, A.L., Coorssen, J.R., Backlund, P.S., Blank, P.S., Humphrey, G.A., Zimmerberg, J., Campbell, J.M. and Vestal, M.L.: De Novo Sequencing of Peptides using MALDI/TOF-TOF. Journal of the American Society for Mass Spectrometry, 13(7), 784-791, 2002

[33] Pitzer, P., Masselot, A. and Colinge, J.: Assessing Peptide De Novo Sequencing Algorithms Performance on Large and Diverse Datasets. Proteomics, 7(17), 3051-3054, 2007

[34] Chen, T., Lu, B.: Algorithms for De Novo Peptide Sequencing using Tandem Mass Spectrometry. Drug Discovery Today: Biosilico, 2(2), 85-90, 2004

[35] Xu, C. and Ma, B.: Software for Computational Peptide Identification from MS-MS Data. Drug Discovery Today, 11(13), 595-600, 2006

[36] Hughes, C., Ma, B. and Lajoie, G.: De Novo Sequencing Methods in Proteomics. Proteome Bioinformatics, 105-121. Humana Press, 2010

[37] Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C. and Zhang, X.: Performance Evaluation of Existing De Novo Sequencing Algorithms. Journal of Proteome Research, 5(11), 3018-3028, 2006

[38] Ma, B.: Challenges in Computational Analysis of Mass Spectrometry Data for Proteomics. Journal of Computer Science and Technology, 25(1), 2010

[39] Paizs, B. and Suhai, S.: Fragmentation Pathways of Protonated Peptides. Mass Spectrometry Reviews, 24(4), 508-548, 2005

[40] Picotti, P., Aebersold, R. and Domon, B.: The Implications of Proteolytic Background for Shotgun Proteomics. Molecular and Cellular Protoemics, 6(9), 1589-1598, 2007

[41] Gerber, F., Krummen, M., Potgeter, H., Roth, A., Siffrin, C. and Spoendlin, C.: Practical Aspects of Fast Reversed-Phase High-Performance Liquid Chromatography Using 3μ Particle Packed Columns and Monolithic Columns in Pharmaceutical Development and Production Working Under Current Good Manufacturing Practice. Journal of Chromatography A, 1036(2), 127-133, 2004

[42] Mann, M., Hendrickson, R.C. and Pandey, A.: Analysis of Proteins and Proteomes by Mass Spectrometry. Annual Review of Biochemistry, 70, 437-473, 2001

[43] Hoopmann, M.R. and Finney, G.L.: High-speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Datasets using High-resolution Mass Spectrometry. Analytical Chemistry, 19(15), 5620-5632, 2007

[44] Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N.G. and Old, W.M.: Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-scale Proteomics Studies. Journal of Proteome Research, 9(8), 4152-4160, 2010

[45] Alves, G., Ogurtsov, A.Y., Kwok, S., Wells, W.W., Wang, G., Shen, R. and Yu, Y.: Detection of Co-eluting Peptides using Database Searching Methods. Biology Direct, 3, 1-16, 2008

[46] Michalski, A., Cox, J. and Mann, M.: More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-dependent LC-MS/MS. Journal of Proteome Research, 10(4), 1785-1793, 2011

[47] Venable, J.D., Dong, M.Q., Wohlschlegel, J., Dilin, A. and Yates, J.R.: Automated Approach for Quantitative Analysis of Complex Peptide Mixture from Tandem Mass Spectra. Nature Methods, 1(1), 39-45, 2004

[48] Hakansson, K., Chalmers, M.J., Quinn, J.P., McFarland, M.A., Hendrickson, C.L. and Marshall, A.G.: Combined Electron Capture and Infrared Multiphoton Dissociation for Multistage MS/MS in a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer. Analytical Chemistry, 75(13), 3256-3262, 2003

[49] Collins, B.C., Gillet, L.C., Rosenberger, G., Rost, H.L., Vichalkovski, A., Gstaiger, M. and Aebersold, R.: Quantifying Protein Interaction Dynamics by SWATH Mass Spectrometry: Application to the 14-3-3 System. Nature Methods, 10(12), 1246-1253, 2013

[50] Liu, Y., Ma, B., Zhang, K. and Lajoie, G.: An Effective Algorithm for Peptide De Novo Sequencing from Mixture Spectra. Proceedings, the 10th International Symposium on Bioinformatics Research and Applications(ISBRA2014), Zhangjiajie, China, June 28-30, 126-137, 2014

[51] Liu, Y., Ma, B., Zhang, K. and Lajoie, G.: An Approach for Peptide Identification by De novo Sequencing of Mixture Spectra. DOI: 10.1109/TCBB.2015.2407401. IEEE Transactions on Computational Biology and Bioinformatics. Accepted for Publication, 2015

[52] Liu, Y., Sun, W., Ma, B., Lajoie, G. and Zhang, K.: An Approach for Matching Mixture MS/MS Spectra with a Pair of Peptide Sequences in a Protein Database, Pro-

ceedings, the 11th International Symposium on Bioinformatics Research and Applications(ISBRA2015), Norfolk, Virginia, USA, June 7-10, 223,234, 2015

[53] Wikipedia Amino Acid Entry: http://en.wikipedia.org/wiki/Amino_acid

[54] The University of Queensland: http://www.di.uq.edu.au/sparqproteins

[55] Pauling L. The Nature of the Chemical Bound, Third Edition, Cornell University Press, 1980

[56] http://www.scienceprofonline.com/chemistry/what-is-organic-chemistry-carbohydrates-proteins-lipids-nucleic-acids.html

[57] Bruce, A., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walters, P.: The Shape and Structure of Proteins. Molecular Biology of the Cell, Fourth Edition, New York and London, Garland Science, 2002

[58] Duncan, M.W., Aebersold, R. and Caprioli, R.M.: The Pros and Cons of Peptide-centric Proteomics. Nature Biotechnology, 28(7), 659-664, 2010

[59] Baenziger, J.U.: A Major Step on the Road to Understanding a Unique Posttranslational Modification and Its Role in a Genetic Disease. Cell, 113, 421-422, 2003

[60] Witze, E.S., Old, W.M., Resing, K.A. and Ahn, N.G.: Mapping Protein Post-Translational Modifications with Mass Spectrometry. Nature Methods, 4(10), 798-806, 2007

[61] Wold, F.: In Vivo Chemical Modification of Proteins (Post-Translational Modification). Annual Review of Biochemistry. 50(1), 783-814, 1981

[62] Oki, M., Aihara, H. and Ito, T.: Role of Histone Phosphorylation in Chromatin Dynamics and Its Implications in Disease. Subcellular Biochemistry, 41, 319-336, 2007

[63] Creasy, D.M. and Cottrell, J.S.: Unimod: Protein Modifications for Mass Spectrometry. Proteomics, 4(6), 1534-1536, 2004

[64] ABRF Delta Mass Database: http://www.abrf.org/index.cfm/dm.home

[65] Mann, M. and Jensen, O.: Proteomic Analysis of Post-Translational Modifications. Nature Biotechnology. 21(3), 255-261, 2003

[66] Han, X., He, L., Xin, L., Shan, B. and Ma, B.: PeakPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications. Journal of Proteome Research, 10(7), 2930-2936, 2011

[67] Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C. and Yates, J.R.: Protein Analysis by Shotgun/Bottom-up Proteomics. Chemical Reviews, 113(4), 2343-2394, 2013

[68] Gustafsson, J.O., Oehler, M.K., Ruszkiewicz, A., McColl, S.R. and Hoffmann, P.: MALDI Imaging Mass Spectrometry(MALDI-IMS)- Application of Spatial Proteomics for Ovarian Cancer Classification and Diagnosis. International Journal of Molecular Sciences, 12(1), 773-794, 2011

[69] Karas, M. and Krger, R.: Ion Formation in MALDI: The Cluster Ionization Mechanism. Chemical Reviews, 103(2), 427-440, 2003

[70] Karas, M., Bachmann, D., Hillenkamp, F.: Influence of the Wavelength in High-Irradiance Ultraviolet Laser Desorption Mass Spectrometry of Organic Molecules. Analytical Chemistry, 57(14), 2935-2943, 1985

[71] Pozniak, B.P. and Cole, R.B.: Current Measurements within the Electrospray Emitter. Journal of American Society of Mass Spectrometry, 18(4), 737-748, 2007

[72] Banerjee, S. and Mazumdar, S.: Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. International Journal of Analytical Chemistry, Article ID: 282574, 2012.

[73] Dawson, P.H.: Quadrupole Mass Analyzers: Performance, Design and Some Recent Applications. Mass Spectrometry Reviews, 5(1), 1-37, 1986

[74] Jonscher, K.R. and Yates, J.R.: The Quadrupole Ion Trap Mass Spectrometer-a Small Solution to a Big Challenge. Analytical Biochemistry. 244(1), 1-15, 1997

[75] Hager, J.W.: A New Linear Ion Trap Mass Spectrometer. Rapid Communication on Mass Spectrometery. 16(6), 512-526, 2002

[76] Mamyrin, B.A.: Time-of-Flight Mass Spectrometry (Concepts, Achievements, and Prospects). International Journal of Mass Spectrometry. 206(3), 251-266, 2001

[77] Marshall, A.G., Hendrickson, C.L. and Jackson, G.S.: Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: a Primer. Mass Spectrometry Reviews, 17(1), 1-35, 1998

[78] Hu,Q., Robert, J.N., Li, H., Makarov, A., Hardman, M. and Cooks, R.G.: The Orbitrap: a New Mass Spectrometer. Journal of Mass Spectrometry. 40(4), 430-443, 2005

[79] Pare, J.J. and Yaylayan, V.: Mass Spectrometry: Principles and Applciations. Techniques and Instrumentation in Analytical Chemistry, 18, 239-266, 1997

[80] Barner-Kowollik, C., Gruendling, T., Falkenhagen, J. and Weidner, S.: Mass Spectrometry in Polymer Chemistry. John Wiley & Sons, 2011

[81] Siuzdak, G.: The Expanding Role of Mass Spectrometry in Biotechnology. MCC Press, San Diego, 2006

[82] Benedikt, J., Hecimovic, A., Ellerweg, D. and Von Keudell A.: Quadrupole Mass Spectrometry of Reactive Plasmas. Journal of Physics D: Applied Physics, 45(40), 403001, 2012

[83] Lin, H.: Algorithms for Characterizing Peptides and Glycopeptides with Mass Spectrometery. PhD Dissertation, University of Waterloo, 2013

[84] Mann, M., Meng, C.K. and Fenn, J.B.: Interpreting Mass Spectra of Multiply Charged Ions. Analytical Chemistry, 61(15), 1702-1708, 1989

[85] Zhang, H.: A New Algorithm for Charge State Deconvolution of Electrospray Ionization Mass Spectra. Master of Science Thesis, University of Western Ontario, 2005

[86] Zhang, Z. and Marshall, A.G.: A Universal Algorithm for Fast and Automated Charge State Deconvolution of Electrospray Mass-to-Charge Ratio Spectra. Journal of the American Society for Mass Spectrometry, 9(3), 225-233, 1998

[87] Zhang, Z., Guan, S. and Marshall, A.G.: Enhancement of the Effective Resolution of Mass Spectra of High-Mass Biomolecules by Maximum Entropy-based Deconvolution to Eliminate the Isotopic Natural Abundance Distribution. Journal of the American Society for Mass Spectrometry. 8(6), 659-670, 1997

[88] Anderson, N.L. and Anderson, N.G.: Proteome and Proteomics: New Technologies, New Concepts, and New Words. Electrophoresis, 19(11), 1853-1861, 1998

[89] McDonald, W.H. and Yates, J.R.: Shotgun Protoemics: Integrating Technologies to Answer Biological Questions. Current Opinion in Molecular Therapeutics, 5(3), 302, 2003

[90] Washburn,W.P. Wolters, D.A. and Yates, J.R.: Large-scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. Nature Biotechnology. 19(3), 242-247, 2001

[91] Wolters D.A., Washburn M.P. and Yates, J.R.: An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. Analytical Chemistry, 73(23), 5683-5690, 2001

[92] Fournier, M.L., Gilmore, J.M., Martin-Brown S.A. and Washburn M.P.: Multidimensional Separations-based Shotgun Proteomics. Chemical Reviews, 107(8), 3654-3686, 2007

[93] Ham, B.M.: Proteomics of Biological Systems: Protein Phosphorylation Using Mass Spectrometry Techniques, pp. 36-37 John Wiley & Sons, 2011

[94] Cooks, R.G.: Collision-Induced Dissociation: Readings and Commentary. Journal of Mass Spectrometry, 30(9), 1215-1221, 1995

[95] Olsen, J.V., Macek, B., Lange, O., Makarov, A., Horning, S. and Mann, M.: Higher-energy C-trap Dissociation for Peptide Modification Analysis, Nature Methods, 4(9), 709-712, 2007

[96] Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J. and Hunt, D.F.: Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. Proceedings of the National Academy of Sciences, 101(26), 9528-9533, 2004

[97] McLafferty, F., Horn, D.M., Breuker, K., Ge, Y., Lewis, M.A., Cerda, B., Zubarev, R.A., Carpenter, B.K.: Electron Capture Dissociation of Gaseous Multiply Charged Ions by Fourier-Transform Ion Cyclotron Resonance. Journal of American Society for Mass Spectrometry, 12(3), 245-253, 2001

[98] Zubarev, R.A., Horn, D.M., Fridriksson, E.K., Kelleher, N.L, Kruger, N.A., Lewis, M.A., Carpenter, B.K. and Mclaferty, F.W.: Electron Capture Dissociation for Structural Characterization of Multiply Charged Protein Cations, Analytical Chemistry, 72(3), 563-573, 2000

[99] Chalkley, R.J.: When Target-decoy False Discovery Rate Estimations are Inaccurate and How to Spot Instances. Journal of Proteome Research, 12(2), 1062-1064, 2003

[100] Gupta, N., Bandeira, N., Keich, U. and Pevzner, P.A.: Target-decoy Approach and False Discovery Rate: When Things May Go Wrong. Journal of the American Society for Mass Spectrometry, 22(7), 1111-1120, 2011

[101] Moore, R.E., Young,M.K., Lee, T.D.: Qscore: An Algorithm for Evaluating SEQUEST Database Search Results. Journal of the American Society for Mass Spectrometry, 13(4), 378-386, 2002

[102] Elias, J.E. and Gygi, S.P.: Target-decoy Search Strategy for Increased Confidence in Large-scale Protein Identificatiion by Mass Spectrometry. Nature Methods, 4(3), 207-214, 2007

[103] Hopper, S., Johnson, R.S., Vath, J.E. and Biemann, K.: Glutaredoxin from Rabit Bone Marrow. Pruification, Characterization, and Amino Acid Sequence Determined by Tandem Mass Spectrometry. Journal of Biological Chemistry, 264(34), 20438-20447, 1989

[104] Bandeira, N., Tang, H., Bafna, V. and Pevzner, P.: Shotgun Protein Sequencing by Tandem Mass Spectra Assembly. Analytical Chemistry, 76(24), 7221-7233, 2004

[105] Bandeira, N., Clauser, K.R. and Pevzner, P.: Shotgun Protein Sequencing: Assembly of Peptide Tandem Mass Spectra from Mixtures of Modified Proteins. Molecular & Cellular Proteomics, 6(7), 1123-1134, 2007

[106] Liu, X., Han, Y., Yuen, D. and Ma, B.: Automated Protein (re)Sequencing with MS/MS and a Homologous Database Yields Almost Full Coverage and Accuracy. Bioinformatics, 25(17), 2174-2180, 2009

[107] Datta, R. and Bern, M.W.: Spectrum Fusion: Using Multiple Mass Spectra for De Novo Peptide Sequencing. Journal of Computational Biology, 16(8), 1169-1182, 2009

[108] Savitski, M.M., Nielsen, M.L., Kjeldsen, F. and Zubarev, R.A.: Proteomics-grade De Novo Sequencing Approach. Journal of Proteome Research, 4(6), 2348-2354, 2005

[109] Bertsch, A., Leinenbach, A., Pervukhin, A., Lubeck, M., Hartmer, R., Baessmann, C., Elnakady, Y.A., Muller, R., Bocker, S., Huber C.G. and Kohlbacher, O.: De Novo Peptide Sequencing by Tandem MS Using Complementary CID and Electron Transfer Dissociation. Electrophoresis, 30(3), 3736-3747, 2009

[110] He, L. and Ma, B.: ADEPTS: Advance Peptide De Novo Sequencing with a Pair of Tandem Mass Spectra. Journal of Bioinformatics and Computational Biology, 8(6), 981-942, 2010

[111] Yan, Y., Kusalik, A.J. and Wu, F.X.: NovoPair: De Novo Peptide Sequencing for Tandem Mass Spectra Pair. In Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on, 150-155, IEEE, 2014

[112] Masselon, C., Pasa-Tolic, L., Li, L.J., Anderson, G.A., Harkewicz, R. and Smith, R.D.: Identification of Tryptic Peptides from Large Database using Multiplexed Tandem Mass Spectrometry: Simulations and Experimental Results. Proteomics, 3(7), 1279-1286, 2003

[113] Zhang, N., Li, X.J., Ye, M.L., Pan, S., Schwikowski, B. and Aebersold, R.: ProbIDtree: An Automated Software Program Capable of Identifying Multiple Peptides from a Single Collision Induced Dissociation Spectrum Collected by a Tandem Mass Spectrometer. Proteomics, 5(16), 4096-4106, 2005

[114] Wang, J., Perez-Santiago, J., Katz, J.E., Mallick, P. and Bandeira, N.: Peptide Identification from Mixture Tandem Mass Spectra. Molecular and Cellular Proteomics, 9(7), 1476-1485, 2010

[115] Wang, J., Bourne, P.E. and Bandeira, N.: Peptide Identification by Database Search of Mixture Tandem Mass Spectra. Molecular and Cellular Proteomics, 10(12), 1476-1485, 2011

[116] Zhang, B., Pirmoradian, M., Chernobrovkin, A. and Zubarev, R.A.: DeMix Workflow for Efficient Identification of Cofragmented Peptides in High Resolution Data-Dependent Tandem Mass Spectrometry. Molecular & Cellular Proteomics, 13(11), 3211-3223, 2014

[117] Ning, K., Fermin, D. and Nesvizhskii, A.I.: Computational Analysis of Unassigned High-Quality MS/MS Spectra in Proteomic Datasets. Proteomics, 10(14), 2712-2718, 2010

[118] Niu, M., Mao, X., Ying, W., Qin, W., Zhang, Y. and Qian, X.: Determination of Monoisotopic Masses of Chimera Spectra from High-resolution Mass Spectrometers Data by Use of Isotopic Peak Intensity Ratio Modeling. Rapid Communications in Mass Spectrometry, 26(16), 1875-1886, 2012

[119] Liu, C., Wang, H.P., Sun, R.X., Fu, Y., Zhang, J.F., Wang, L.H., Chi, H., Li, Y., Xiu, L.Y., Wang, W.P. and He, S.M.: pParse: A Method for Accurate Determination of Monoisotopic Peaks in High-resolution Mass Spectra. Proteomics, 12(2), 226-235, 2010

[120] Luethy, R., Kessner, D.E., Katz, J.E., MacLean, B., Grothe, R., Kani, K., Faca, V., Pitteri, S., Hanash, S., Agus, D.B. and Mallick, P.: Precursor-Ion Mass Re-Estimation Improves Peptide Identification on Hybrid Instruments. Journal of Proteome Research, 7(9), 4031-4039, 2008

[121] Frank, A.M., Savitski, M.M., Nielsen, M.L., Zubarev, A. and Pevzner, P.A.: De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry. Journal of Proteome Research, 6(1), 114-123, 2007

[122] Snyder, A.P.: Interpreting Protein Mass Spectra: A Comprehensive Resource. London, U.K.: Oxford University Press, 2000

[123] Vizcaino, J.A., Cote, R.G., Csordas, A., Dianes, J.A., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R. and Hermjakob, H.: The Proteomics Identification(PRIDE) Database and Associated Tools: Status in 2003. Nucleic Acid Research, 41(D1), D1063-D1069, 2013

[124] Henry, L., Deutsch, E.W. and Aebersold, R.: Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics. Journal of Proteome Research, 9(1), 605-610, 2009

Appendix A

Annotation of Identifications

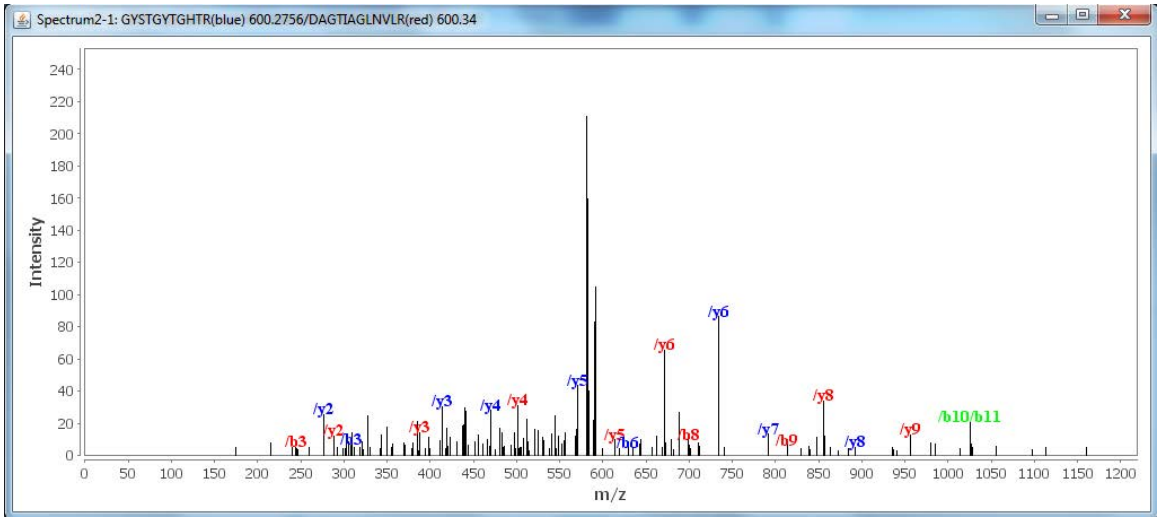

Figure A.1: For this input mixture spectrum, the best matched pair are **GYSTGYTGHTR** and **DAGTIAGLNVLR**.

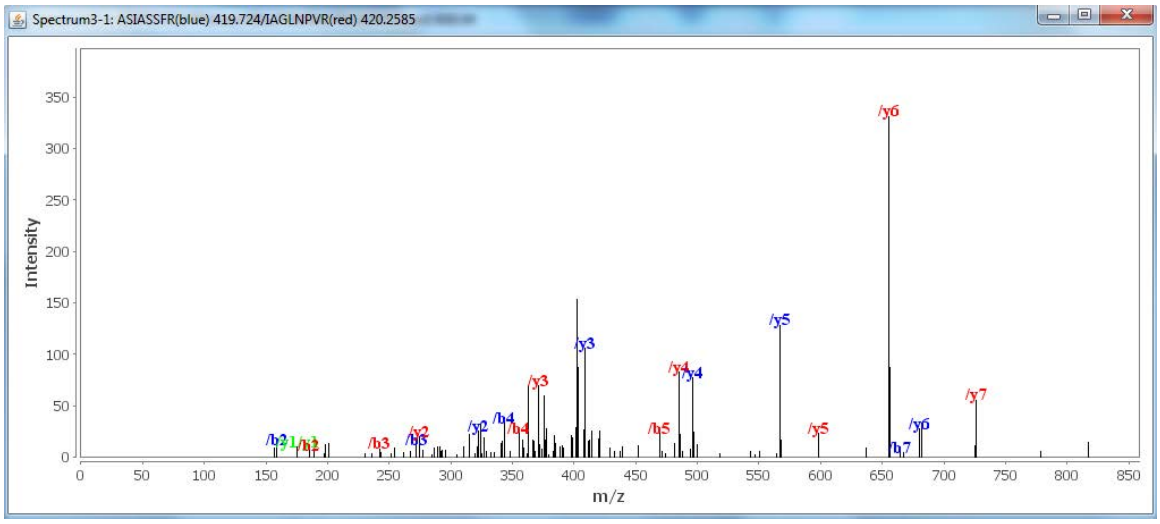

Figure A.2: For this input mixture spectrum, the best matched pair are **ASIASSFR** and **IAGLNPVR**.

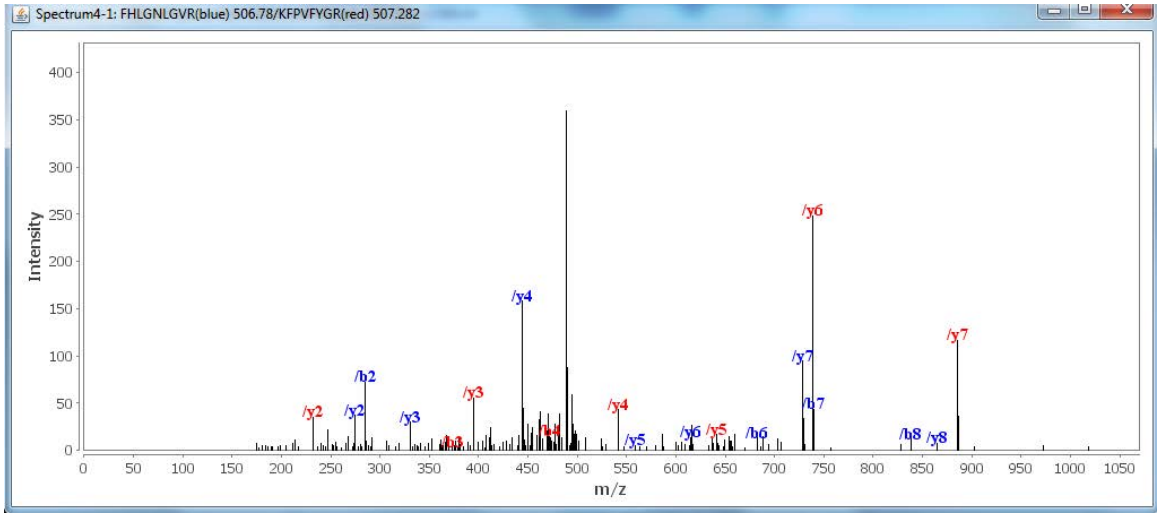

Figure A.3: For this input mixture spectrum, the best matched pair are **FHLGNLGVR** and **KFPVFYGR**.

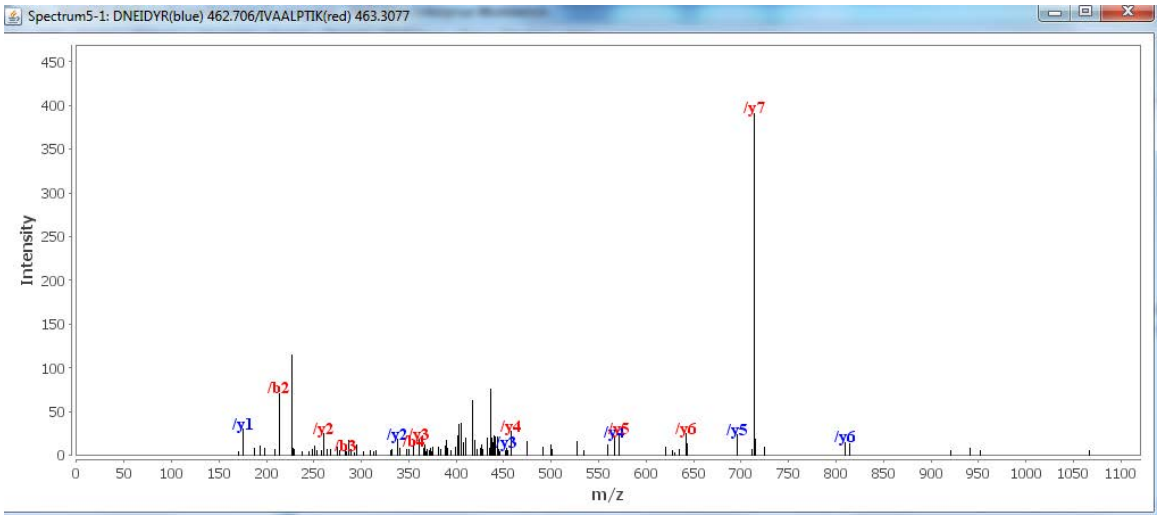

Figure A.4: For this input mixture spectrum, the best matched pair are **DNEIDYR** and **IVAALPTIK**.

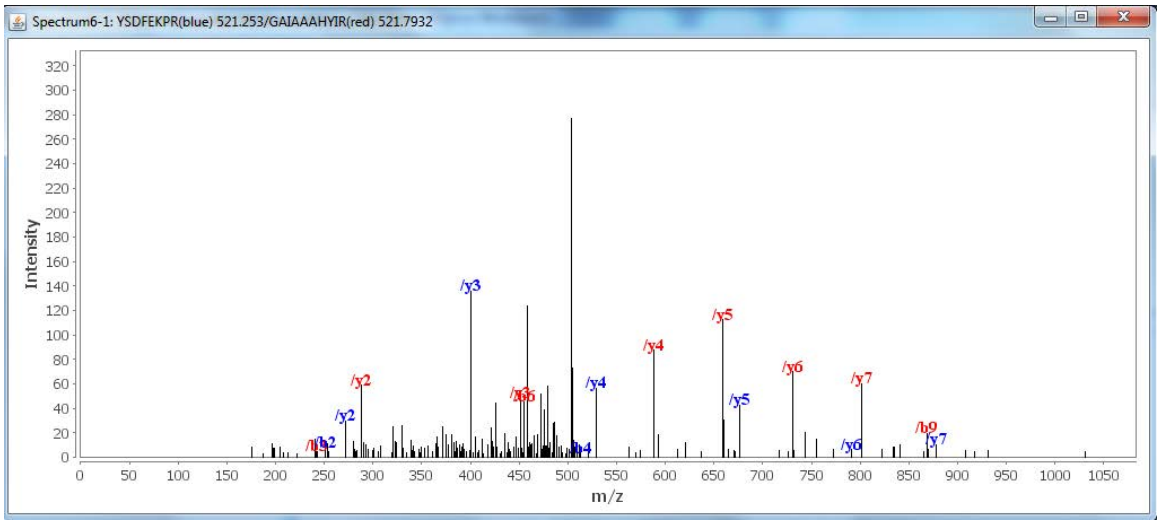

Figure A.5: For this input mixture spectrum, the best matched pair are **YSDFEKPR** and **GAIAAAHYIR**.

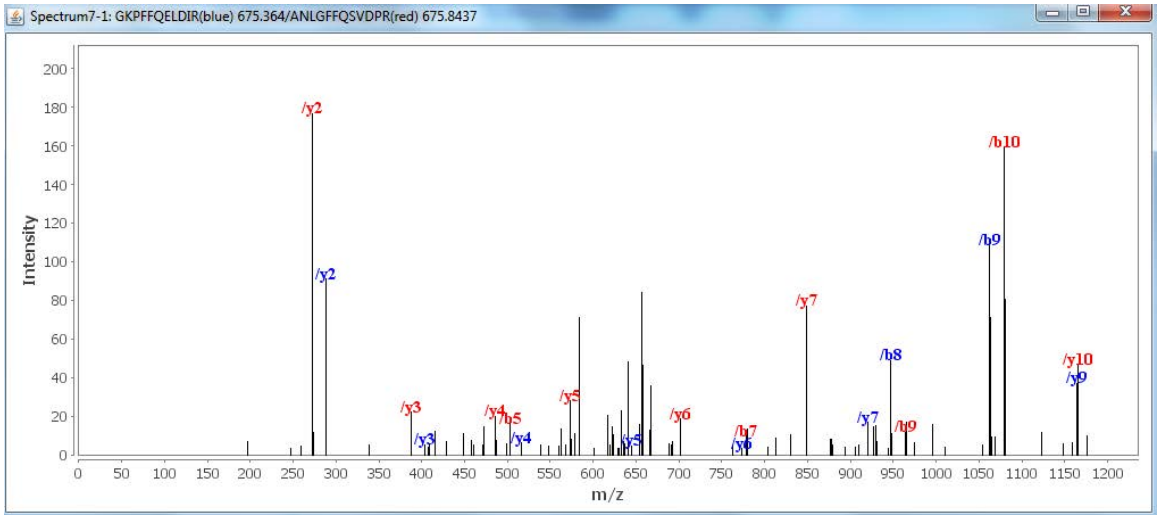

Figure A.6: For this input mixture spectrum, the best matched pair are **GKPFFQELDIR** and **ANLGFFQSVDPR**.

Curriculum Vitae

Name:	Yi Liu
Post-Secondary Education and Degrees:	Central South University Changsha, Hunan Province, P.R. China 2004-2008 B. Eng., Computer Science and Technology Central South University Changsha, Hunan Province, P.R. China 2008-2011 M. Eng., Computer Science and Technology
Honours and Awards:	Western Graduate Research Scholarship 2011-2015
Academic Training	Teaching Assistant The University of Western Ontario 2011-2015 Research Assistant The University of Western Ontario 2011-2015

Publications:

1. Liu, Y., Ma, B., Zhang, K. and Lajoie, G.: An Approach for Peptide Identification by De novo Sequencing of Mixture Spectra. DOI: 10.1109/TCBB.2015.2407401. IEEE Transactions on Computational Biology and Bioinformatics. Accepted for Publication, 2015

2. Liu, Y., Ma, B., Zhang, K. and Lajoie, G.: An Effective Algorithm for Peptide De Novo Sequencing from Mixture Spectra. Proceedings, the 10th International Symposium on Bioinformatics Research and Applications (ISBRA2014), Zhangjiajie, China, June 28-

30, 126-137, 2014

3. Liu, Y., Sun, W., Lajoie, G., Ma, B., and Zhang, K.: An Approach for Matching Mixture MS/MS Spectra with a Pair of Peptide Sequences in a Protein Database. Proceedings, the 11th International Symposium on Bioinformatics Research and Applications (ISBRA2015), Norfolk, Virginia, USA, June 7-10, 223-234, 2015

4. Liu, Y., Sun, W., Lajoie, G., Ma, B., and Zhang, K.: De Novo Sequencing Assisted Approach for Characterizing Mixture MS/MS Spectra. (Invited and Submitted to IEEE Transactions on Nanobioscience)