Mass Spectrometry Based De Novo Peptide Sequencing Error Correction

by

Chenyu Yao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer science

Waterloo, Ontario, Canada, 2017

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Extensive study has been conducted on the identification of peptide sequences with mass spectrometry. With the development of computer hardware and algorithms, de novo sequencing has drawn attention from researchers for many years. Because it does not require a protein database, de novo sequencing is able to serve as either a complement of database searching or a stand alone method. As shown by Novor [1], the speed of de novo sequencing significantly exceeds the speed of protein database searching. Improving the accuracy of de novo sequencing is essential.

Overlapping peptides occur quite frequently in a typical heavy chain proteomics sample. In this thesis, we have proposed an algorithm to efficiently and reliably detect the overlapping peptides. In addition, two strategies named labeling and voting are designed to utilize overlapping peptides so as to improve the accuracy of de novo sequencing.

According to the results, the effect of our labeling strategy is not obvious with the current version of Novor. Although the improvemnt made by the labeling strategy is not significant, we still demonstrate the potential of the method. However, the performance of our voting strategy is surprising and noteworthy. It is able to achieve significant improvement of de novo sequencing with little running time.

Acknowledgements

I would like to thank all who made this thesis possible.

I would first like to thank my thesis supervisor Prof. Bin Ma. The door to Prof. Ma's office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this thesis to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to thank the rest of my thesis committee: Prof. Forbes Burkowski and Prof. Kaizhong Zhang for their insightful comments and questions.

I thank my fellow labmates in the Biofomatics Group: Jianqiao Shen, Lian Yang, Rong Wang, Qi Tang, and Tiancong Wang, for the heated discussions, and for all the fun we have had in the last two years.

Last but not least, I would like to thank my family for supporting me spiritually throughout my life. I have a special feeling of gratitude to my lovingparents, Qiuxing and Xiadong whose words of encouragement and push for tenacity ring in my ears.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In proteomics, identifying protein or peptide sequences is a common task. In a typical experiment, protein samples are digested with enzymes and measured with tandem mass spectrometry. Millions of spectra are generated in such an experiment. Each of these spectra is presumed to come from the measurement of a single peptide and is thus used to derive the peptide's amino acid sequence. Due to the impracticability of manually analyzing such a large amount of data, automatically analyzing those spectra by computer is necessary.

At present, protein database searching and de novo sequencing are the two main computational approaches to identify peptide sequences from mass spectrum data. In protein database searching, each spectrum is compared to the peptides in a protein database in order to find highly confident peptide-spectrum matches (PSMs). Protein database searching has been studied for many years by a large number of researchers. When a protein database is not available, de novo sequencing is an alternate choice. The de novo sequencing focuses on directly deriving sequences from the mass spectrum. Compared with protein database searching, the accuracy and efficiency of de novo sequencing still have much room for improvement.

With the development of Novor[1], the speed of de novo sequencing may exceed protein database searching. Improving the accuracy of de novo sequencing is now urgent and therefore is the main objective of this thesis.

1.2 Research Objectives and Contributions

In proteomics, we frequently observe multiple peptides sharing a common substring. In this thesis, we refer to these as overlapping peptides. In theory, there are a total of four cases of overlapping peptides as shown in Figure 1.1. In this thesis, we address two of these four cases in which one peptide's sequence is the prefix or suffix of another peptide's sequence (see (a) and (b) in Figure 1.1). These two cases are the most common because they can occur due to the enzyme non-specific cleavages during sample preparation.

Figure 1.1: Four types of overlapping peptide

If the ms/ms spectra of two overlapping peptides are used together in de novo sequencing, we would be able to improve the overall result. This thesis proposes two methods to improve the de novo result, namely, labeling and voting. These two methods may be used either separately or together.

The novel contribution of this thesis includes: (1) a general algorithm and a rigorous scoring function to detect overlapping peptides in mass spectrum datasets, (2) a method to classify fragment peaks into different ion types using overlapping peptides, and (3) an innovative strategy to correct the errors in de novo sequencing results utilizing overlapping peptides.

1.3 Thesis Overview

1.3.1 Method

The overall procedure for our method consists of three parts: overlapping peptide detection, labeling and voting. Overlapping peptide detection is a pre-requisite for the latter two steps. The labeling and the voting are two parallel strategies that can be used together or separately. The overall structure is shown as Figure 1.2. The functions of these parts are summarized in this section.

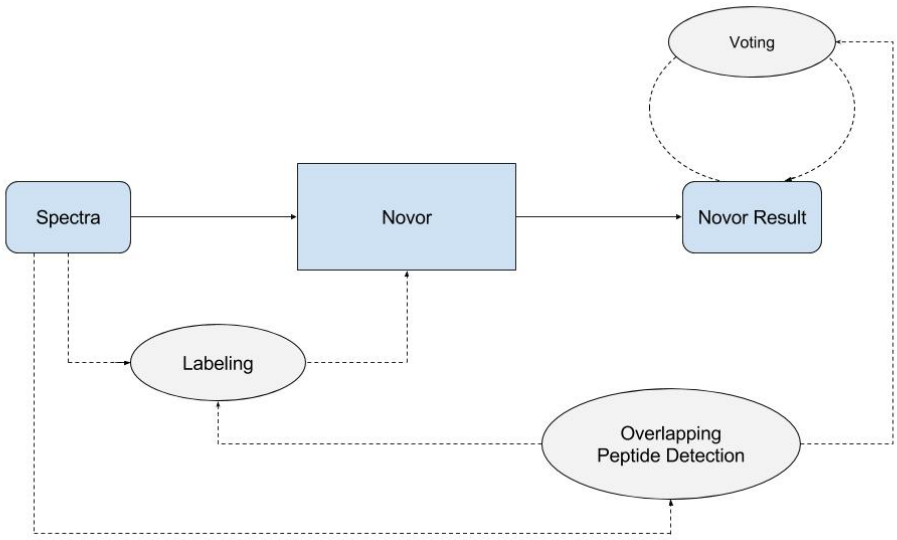


Figure 1.2: Overall structure

Part 1. In the *Overlapping Peptide Detection* part, overlapping peptide pairs are detected. A pairing score is calculated among the candidates and a threshold is used to filter possible pairs.

Part 2. In the *Labeling* part, the original spectrum files are modified. Peaks of the fragment ions containing the peptide's N-terminus and C-terminus, respectively, are separated and labeled. Novor is modified to process the labeled spectrum files and to generate a new result.

Part 3. The *Voting* part modifies the Novor result sequences based on the overlapping peptide pairs. By comparing the amino acid score between two overlapping peptides, the sequences with lower scores are modified.

Since the result of the overlapping peptides detection process directly affects the following two parts, the accuracy and efficiency of the overlapping peptides detection has a significant influence on the overall performance of the method. A threshold must be set to balance the precision and efficiency of the algorithm. We have drawn a precision-recall curve to show the effects of different thresholds.

The labeling method only slightly improves Novor's de novo result. According to the experiments we conducted, the labeling strategy can achieve 1% improvement in terms of correct amino acids for all overlapping peptides. The method also requires Novor to run three times. The possible reasons why this method does not perform as well as we expected are analyzed in Section 4.3.

Finally, the voting method greatly improves the de novo result creating an approximately 3-8% increase in correctly identified amino acids among overlapping peptides. The experiment was conducted on multiple datasets and all of them show significant improvement. Since the overlapping peptide detection part can be run in parallel with Novor and the running time of the voting is fast, we conclude that the voting strategy is practical for de novo sequencing.

1.3.2 Thesis Structure

This thesis consists of this introduction and an additional five chapters:

Chapter 2 introduces the background and related works.

Chapter 3 describes the algorithm and scoring function for overlapping peptide detection.

Chapter 4 introduces the method of separating different fragment ion peaks and the modification of the Novor software to use such separation to improve de novo sequencing.

Chapter 5 describes the correction of de novo sequencing errors by utilizing overlapping peptides.

In addition, experiment results for each of the above methods are presented in the corresponding chapters.

Chapter 6 summarizes the thesis and proposes future work.

Chapter 2

Background and Related Works

2.1 Background

2.1.1 Proteins, Peptides, and Amino Acids

Proteins are complicated biomolecules within an organism. Proteins are the main carrier of any functions an organism performs. They are coded and translated from DNA. They consist of one or more amino-acid chains.

Peptides are linear chains of amino acid residues. Peptides are distinguished from proteins on the basis of size and there are no distinct boundaries between peptides and proteins. Proteins are usually broken down into smaller peptide chains by digestive enzymes. Identifying the sequence of peptides is the main topic of this thesis.

Amino acids are organic compounds that contain an amino, a carboxyl, and a side chain as shown in Figure 2.1. There are 20 different amino acids which differ by side chain. Different side chains have different amino acid masses with the exception of Isoleucine(I) and Leucine(L), which have the same mass. In this thesis, we are more concerned with the residue mass of an amino acid. When amino acids combine to form peptides or proteins, water is removed in this condensation reaction. In this case, the amino acid after the loss of water is called an amino acid residue. Table 2.1 shows the residue mass table of 20 amino acids. The unit of mass is Da which equals 1/12 of the mass of carbon.

The mass of one peptide is equal to the sum of all amino acid residues' masses plus the mass of H_2O. Thus, by measuring the mass of a protein sequence and its substring, it is possible to identify the protein. This process is the basis of protein identification.

Amino Acid	Single Letter Code	Residue Mass
Glycine	G	57.02147
Alanine	A	71.03712
Serine	S	87.03203
Proline	P	97.05277
Valine	V	99.06842
Threonine	T	101.04768
Cysteine	C	103.00919
Isoleucine	I	113.08407
Leucine	L	113.08407
Asparagine	N	114.04293
Aspartic	D	115.02695
Glutamine	Q	128.05858
Lysine	K	128.09497
Glutamic	E	129.04260
Methionine	M	131.04049
Histidine	H	137.05891
Phenylalanine	F	147.06842
Arginine	R	156.10112
Tyrosine	Y	163.06333
Tryptophan	W	186.07932

Table 2.1: Residue mass of amino acids

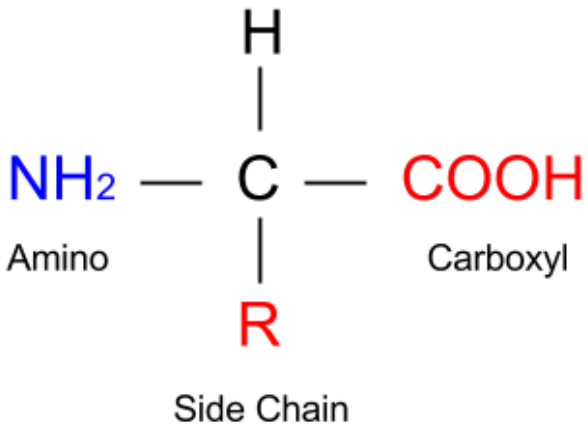


Figure 2.1: Amino Acid Structure

2.1.2 Post-Translational Modifications (PTMs)

PTMs are chemical modifications that play a key role in functional proteomics because they regulate activity, localization, and interaction with other cellular molecules such as proteins, nucleic acids, lipids, and cofactors [2]. PTMs occur in amino acid side chains or peptide linkages. These modifications include phosphorylation, glycosylation, ubiquitination, methylation, acetylation, lipidation and hydroxylation, as shown in Figure 2.2.

The existence of PTMs greatly complicates peptide identification. Hundreds of PTMs have been discovered. Most peptide identification application supports only a few of them. Typically, these applications let users include a small number (less than six) of known PTMs during the set-up procedure for peptide identification. More inclusion of PTMs would exponentially increase the running time of the software.

In this thesis, we intentionally created Cysteine carbamidomethylations to reduce and block the disulphide bonds. Carbamidomethylation is a deliberate post-translational modification introduced to Cysteine (C) residues by reacting with iodoacetamide [4].

Peptides with PTMs must be distinct from overlapping peptides. A filtering procedure is introduced for this purpose as discussed in detail in Section 3.2.2.

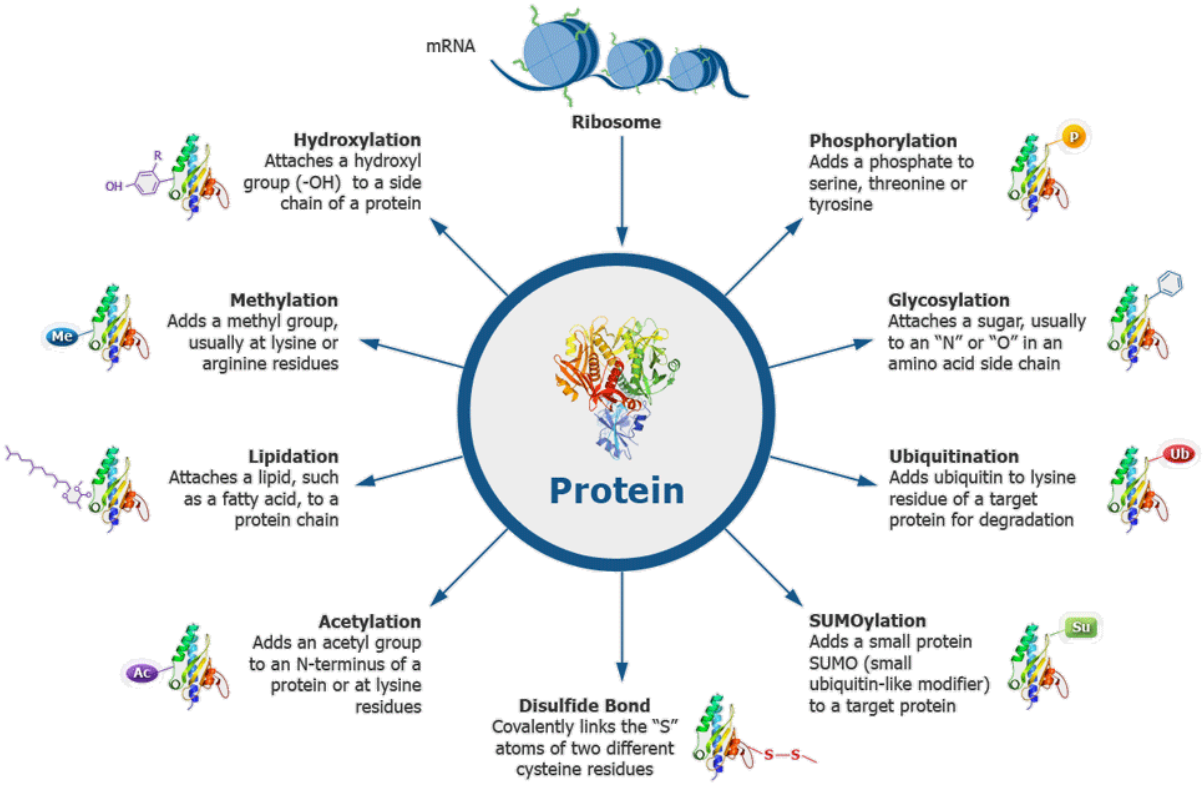


Figure 2.2: Overview of PTMs [3]

2.1.3 Basics of Mass Spectrometry

Mass Spectrometry

Mass spectrometry (MS) is widely used in the fields of chemistry and biology. It has been used since the 1980s to measure the mass of particles. In bioinformatics, mass spectrometry analysis of proteins measures the mass-to-charge ratio (m/z) of ions to detect, identify and quantify molecules in simple and complex mixtures [5].

A mass spectrometer contains an ion source, a mass analyzer, and an ion detector. Samples are loaded into the ion source chamber and then vaporized and ionized. Ions are accelerated because of the charges they receive. The mass analyzer accelerates ions in magnetic fields or electrical fields and ions with different m/z are deflected by different amounts. Thus, the mass analyzer can be used to separate ions for global analysis or to

filter out specific ions for tandem mass spectrometry. The ion detector detects ions passing through the mass analyzer and produces a signal from the separated ions. After the entire process is complete, mass spectra are produced. A mass spectrum is an intensity vs. m/z plot in which each peak represents a signal of the ions.

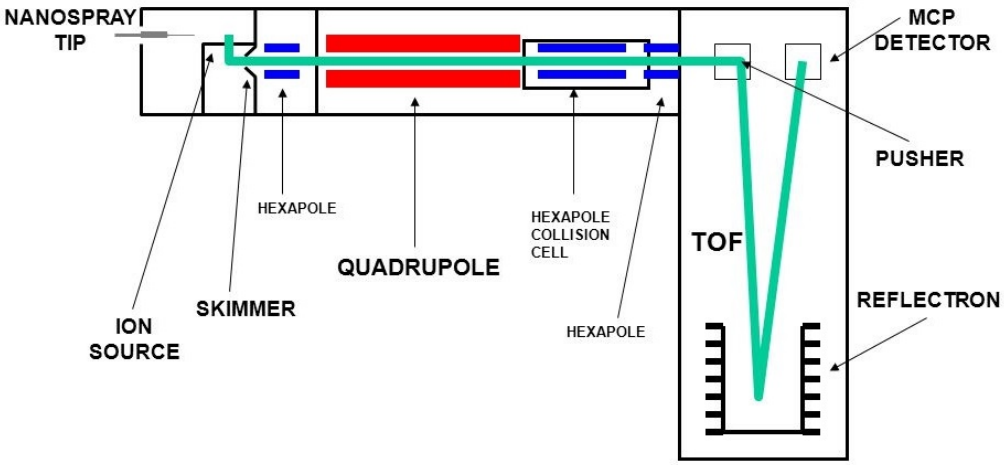


Figure 2.3: Overview of a TOF mass spectrometer [6]

At present, mass spectrometry used in proteomics have different mechanisms, which can be differentiated by the type of ion source or mass analyzer. MALDI and ESI are two different ion sources. Ions coming from MALDI [9] usually contain one positive charge while ions from ESI [8] usually contain one or more charges. The most often used mass analyzers in proteomics are ion trap [10], quadrupole [11], time of flight (TOF) [12], Fourier transform [13] and orbitrap [14]. Figure 2.3 shows the flow chart of a TOF mass spectrometer.

In theory, every signal peak in a spectrum represents an ion from the sample. However, in reality, spectra from the experiment are much more complicated than under theoretical conditions. Isotope peaks, the mass error of the instrument and noise peaks are some types of disturbance that must be taken into consideration in mass spectrometry.

Tandem Mass Spectrometry (MS/MS)

Tandem mass spectrometry (MS/MS) involves multiple steps of mass selection and analysis. Ions of a particular m/z are filtered out in the first stage (MS1). These are called precursor

ions. They are then fragmented by a fragmentation mechanism. The resulting fragment ions are detected in the second stage (MS2). MS2 spectra are the final result. Masses of fragment ions in tandem mass spectrometry provide the structural information of peptides and thus makes peptide sequencing possible. In our research, MS2 spectra are the main study focus.

Collision induced dissociation (CID) [16] is one of the most often used fragmentation mechanisms. In theory, fragmentation can occur in any position of peptide chains and form a, b, c, x, y and z fragment ions, as shown in Figure 2.4. The most significant peaks in CID fragmentation represent b-ions and y-ions. Thus, given the sequence of the peptide and the corresponding spectrum, we are able to calculate the theoretical fragment ion mass and match it with the spectrum. Figure 2.5 shows an example of a peptide and its CID spectrum match. Ideally, the matched peptide sequence meets the following two conditions:

- Most high-intensity peaks can be explained by fragment ions;
- Most fragment ions can be matched with corresponding peaks in spectra.

Therefore, the above two conditions constitute the significant evidence when evaluating the matched spectrum. Figure 2.5 shows an ideal match of peptide and spectra.

However, in reality, spectra from most experiments are not ideal. Since the fragment mechanism is complex, some theoretical fragment ions calculated from such a simple model may not be present in the spectrum, and some peaks in the spectra cannot be easily explained. For example, peptides may be fragmented multiple times, resulting in internal fragment ions. In such cases, peaks in the spectrum are very complicated and hard to identify. The details of peptide identification are introduced in next section.

2.1.4 Peptide Sequence Identification

Peptide sequencing is the technique of determining the amino acid sequence of all or part of a peptide. There are currently two approaches for sequence identification: database searching and de novo sequencing.

Database Searching

Database searching is a common and well-developed peptide identification technique and has long been studied. Common database searching tools include Mascot [18], SEQUEST

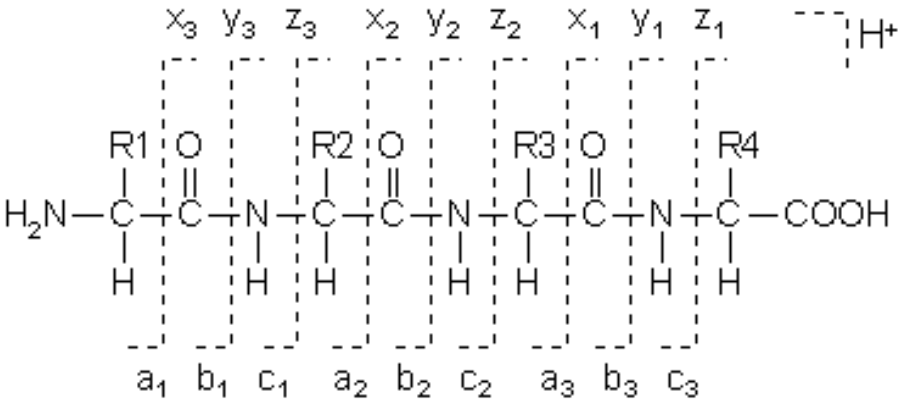


Figure 2.4: Different fragment positions forms different types of ions [15]

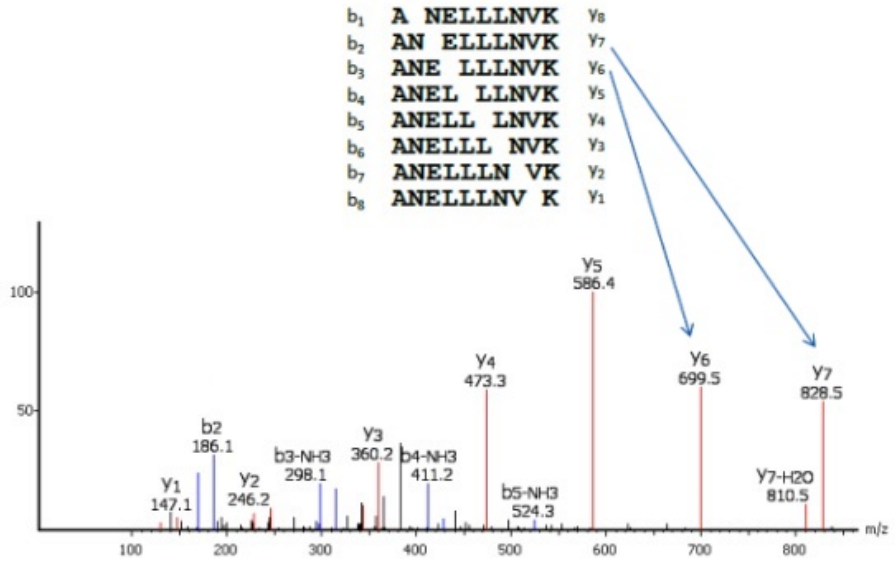


Figure 2.5: Match of a peptide and an experiment spectrum [17]

[19], MaxQuant [20], OMSSA [21], X!Tandem [22], ProteinProspector [23] and MS-GFDB [24].

The basic idea of database searching is to match the spectrum with an existing protein database. The typical process of database searching for one spectrum is as follows. First, all

peptides which have mass within a certain error tolerance of the precursor mass are selected. Then, the theoretical spectra of candidate peptides are compared with an experiment spectrum. A score function is calculated to select the best possible matches. The score function is typically influenced by the current spectrum and the matching conditions of other spectra in the same experiment sample.

Database searching is heavily dependent on the protein database. If a suitable database is available, the accuracy and efficiency of this technique are assured. However, if the protein database is not available, de novo sequencing better serves the role of peptide identification.

De Novo Peptide Sequencing

De novo is a Latin phrase meaning *from the beginning*. Unlike database searching, de novo sequencing attempts to derive peptide sequences directly from spectra. De novo sequencing can be defined as a purely mathematical problem, so it draws attentions from many mathematicians. Typical de novo sequencing software includes Novor, PEAKS [25], PepNovo [26], Lutefisk [27], and NovoHMM [28]. A more comprehensive review of the de novo sequencing software may be found in [29].

In more detail, the task of de novo sequencing can be defined as follows:

Find a peptide sequence that has the total residue mass equal to the given mass and its fragment ions "adequately explain" the peaks in the spectra.

It is important to mathematically define *adequately explain* in the above statement. A scoring function has been introduced to clearly define how well the fragment ions explain the spectra. Dancik et al [30] devised a scoring function for de novo sequencing and variations of their work have been widely used for modern de novo software packages.

De novo sequencing and protein database searching were once considered to be two separate approaches for peptide identification. However, researchers have realized that even if a protein database is available, experiment samples may contain peptides that are not listed in databases. Moreover, de novo sequencing can be used to assist database searching to improve sensitivity and accuracy. The Peaks DB [31] software relies heavily on de novo sequencing results to improve the filtration and the scoring function. This combination results in significantly improved sensitivity and accuracy in comparison to existing database searching software.

De novo sequencing was once considered slow compared with protein databases searching. However, with the release of Novor, speed is no longer a disadvantage of de novo

sequencing. Novor can sequence more than 300 MS/MS spectra per second on a laptop computer. This surpasses the acquisition speed of current mass spectrometers and, therefore, creates a new possibility to perform de novo sequencing in real time while the spectrometer is acquiring the spectral data[1].

In this thesis, we focus on improving the accuracy of Novor. Therefore, a brief introduction of Novor is presented below.

Novor was created on the basis of the decision tree modeling in machine learning. Decision trees with 169 features and thousands of nodes were derived from training data. A new scoring function was designed to evaluate the match between a sequence and its spectrum. An algorithm combining both dynamic programming and heuristics was developed to select the best matches.

Figure 2.6 shows an example of a Novor output file. Novor has output the overall confidence score and amino acid score (aaScore). In this thesis, the labeling method compares the overall confidence score among different cases while the voting strategy makes use of the aaScore for peptide substring substitution.

21	# id	scanNum	RT	mz(data)	z	pepMass(	err(data-	ppm(1e	score	peptide	aaScore
22	1	4	1.2	1127.873	1	1126.828	0.0374	33.2	0.4	VKKKKKKKL	8-1-1-1-1-1-1-1-5
23	2	5	1.3	1473.68	1	1472.651	0.0212	14.4	2.6	SETPYWWWRY	33-1-1-1-1-1-1-1-1-5
24	12	16	2.9	1068.819	1	1067.779	0.0318	29.8	0.5	VVKLKKKLL	8-1-1-1-1-1-1-1-5
25	16	20	3.4	1207.729	1	1206.71	0.0116	9.6	4	LLKWARRQH	35-1-1-1-1-1-1-1-5
26	27	32	5	1127.85	1	1126.828	0.0148	13.1	1.4	KKKKKKKVL	8-1-1-1-1-1-1-1-5
27	65	84	12.8	667.1764	1	666.1625	0.0066	10	1.9	Q(Pyro-Glu)DDC(Ca	1-1-1-1-8

Figure 2.6: Example of a Novor output file

2.2 Related works

Although the spectra of related peptides have been implemented to solve different problems, they have not yet been used for improving the accuracy of de novo sequencing. In this section, we will review two related works.

2.2.1 MS-PSA

Mass spectrometry-peak shift analysis (MS-PSA) [32] was developed by Thomas Wilhelm and Alexandra M. E. Jones to identify post-translational modifications (PTMs).

Currently, most peptide identification software applications only allow users to include a few (typically less than ten) known PTMs in an experiment sample before running the experiment. Despite the fact that hundreds of PTMs have been discovered, a peptide identification algorithm is unable to consider all of these PTMs without increasing the running time to an unacceptable level. In addition, there remain unknown PTMs, which further complicates the peptide identification problem. MS-PSA is not restricted by these obstacles and is complementary to standard protein database searching tools such as MASCOT and SEQUEST.

MS-PSA focuses on related peptides. Two peptides are related if they share the same substring. Unmodified and modified versions of the same peptide are related peptides. The overlapping peptides discussed in this thesis are also considered related peptides.

MS-PSA uses the MS/MS spectrum as input. Spectra are divided into different groups according to their precursor mass and peaks pattern. For each group, common peaks in each spectrum are detected and combined into a "fingerprint". This "fingerprint" is the key feature for detecting related peptides. Two groups of spectra are recognized as related by matching the peaks of their "fingerprints". Potential peak shifts are identified if two spectra are related. PTMs are detected based on peak shift.

2.2.2 Spectral Networks

Nuno Bandeira et al [33] have proposed a strategy to speed up peptide database searching by building spectral networks.

These authors claim that existing approaches that compare spectra with those focused in protein databases have reached a bottleneck. Consequently, they develop a new conceptual approach to protein database searching.

Their method detects related peptides by matching peaks and comparing patterns of candidate spectra. After detecting related peptides, spectral networks are constructed. Using these spectral networks, prefix and suffix laddering peaks are separated, noise peaks are reduced, and peptide reconstructions that may contain the correct one are generated.

The database searching method is performed in an extremely fast pattern-matching algorithm. Instead of comparing spectra with the database separately, spectra are consolidated into clusters. By matching the features of each cluster with the protein database, the speed of peptide identification is greatly accelerated. In addition, PTMs can be detected by spectral networks.

Chapter 3

Overlapping Peptide Pairs Detection

3.1 Overlapping Peptides

In this thesis, our main focuses are overlapping peptides sharing the same N-terminus or C-terminus, as shown in case (a) and case (b) in Figure 1.1. During the sample preparation, peptides are produced either intentionally or unintentionally overlapping. An enzyme, trypsin, was used to digest the protein into shorter peptides. Normally, trypsin digests after amino acids K or R but not before P. However, the enzyme digestion is not always specific. One end (or sometimes both ends) of the resulting peptide may not follow the trypsin digestion rule. The clustering of overlapping peptides shown in the red circle in Figure 3.1 consists of one peptide, FSGSGSGTDRTLK, that follows the digestion rule, and several other peptides that only follow the rule at the C-terminal end.

The goal in this chapter is to detect overlapping peptides in order to increase the de novo sequencing accuracy. The peptide sequences are not available and we would have to detect overlaps from spectra. Thus, we need to discover the special features of overlapping peptides.

Based on observation, if two peptides are overlapping, their corresponding spectra share the most b-ion peaks or y-ion peaks. Figure 3.2 shows the spectra of the sequence VTC(Cam)VVVDISKD and the sequence VTC(Cam)VVVDISK. "Cam" inside parenthesis represents the carbamidomethylation of the preceding Cysteine residue. In the figure, b-ion peaks are highlighted in blue and y-ion peaks are highlighted in black. As we see, the b-ion peaks in both spectra match and y-ion peaks are shifted by approximately 115 Da. Tables 3.1 and 3.2 show the theoretical masses of b-ion peaks and y-ion peaks. From

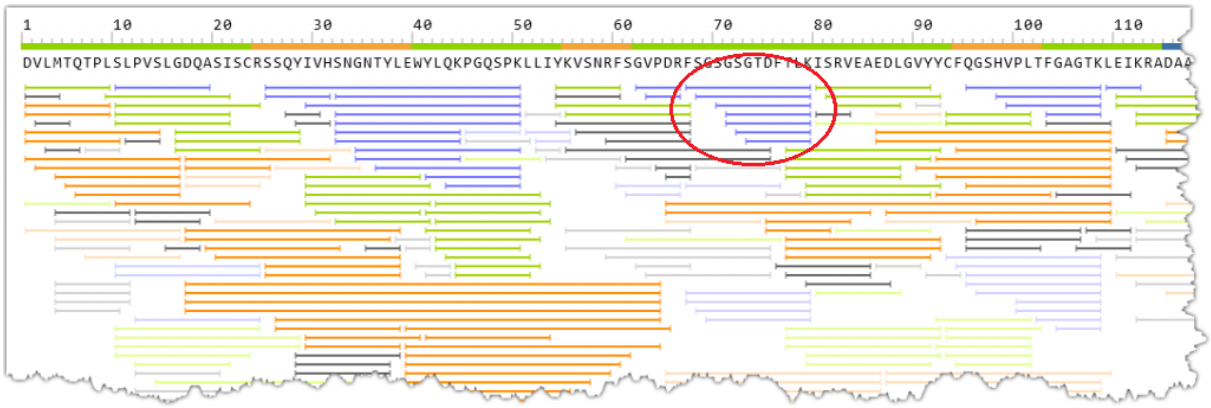


Figure 3.1: Red circle shows an example of overlapping peptides [34]

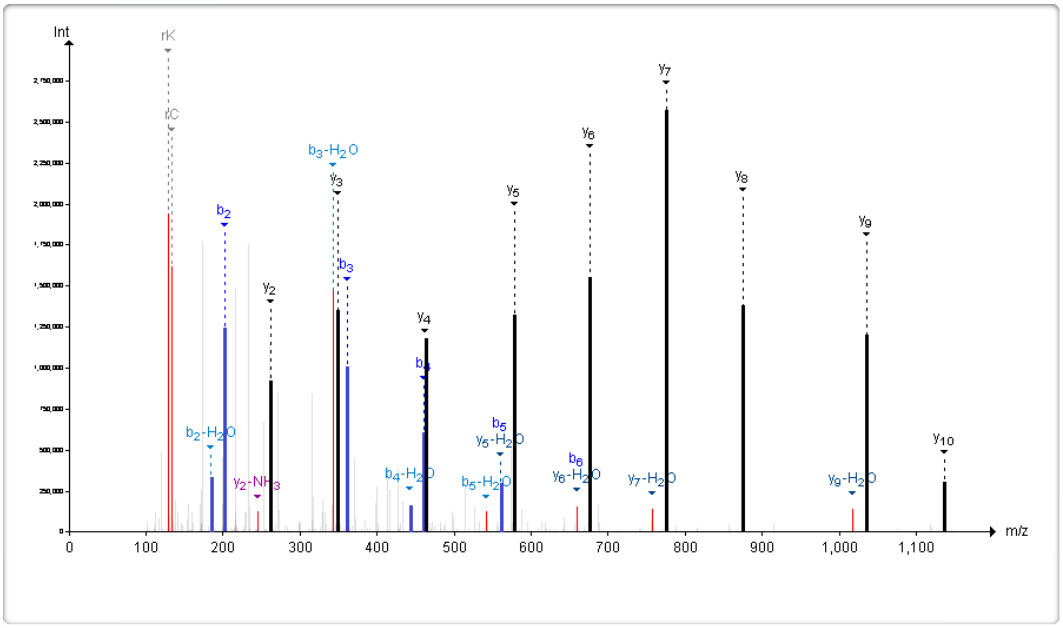
the theoretical masses and spectrum data, we can conclude that all b-ions of sequence VTC(Cam)VVVDISK match the b-ions (b_1 to b_9) of sequence VTC(Cam)VVVDISKD. Furthermore, the differences between all y-ions of sequence VTC(Cam)VVVDISK and the corresponding y-ions (y_2 to y_{10}) of sequence VTC(Cam)VVVDISKD are the same and are equal to the mass difference of their corresponding precursor (115.02694 Da). The fifth column of Table 3.2 illustrates this phenomenon.

For ease of description, we define two terms as follows. When comparing two spectra, we define *matching peaks* to be those have same m/z in both spectra. In addition, we define *shifted peaks* as follows: suppose spectrum s_1 contains peak p_1 and spectrum s_2 contains peak p_2, if the difference between p_1 and p_2 is exactly equal to the precursor mass difference of s_1 and s_2, then we define p_1 and p_2 to be shifted peaks. If two peptides overlap and share the same N-terminus, then their shared b-ion peaks are matching peaks and most of their y-ion peaks are shifted peaks, and vice versa.

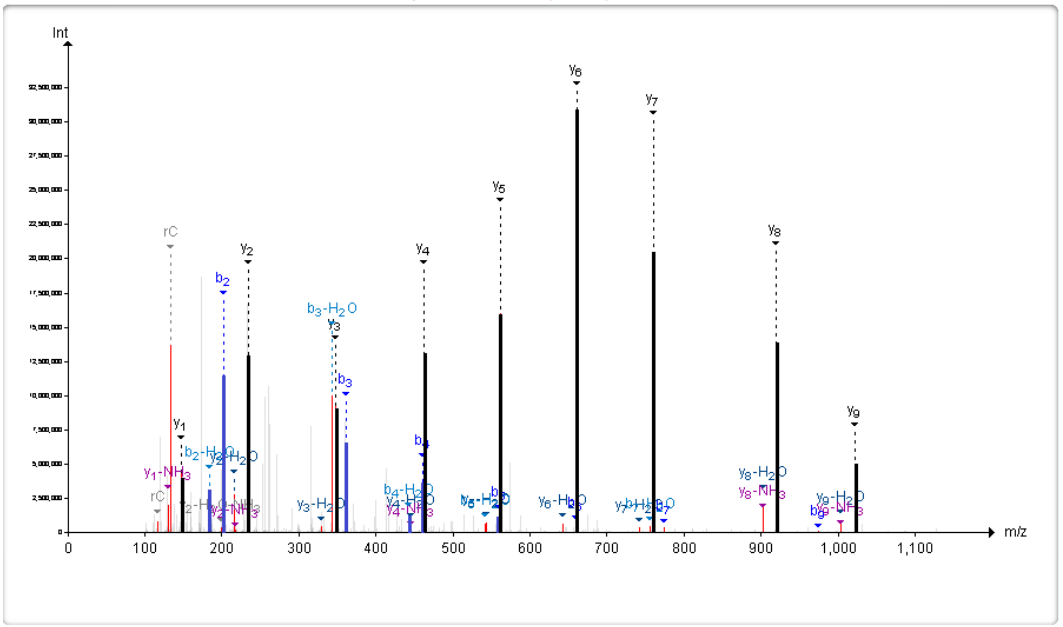
Making use of the above observations is the main approach to identify overlapping peptides. By searching matching peaks and shifted peaks in spectra and evaluating their quality, we are able to detect overlapping peptides.

3.2 Methodology

This section describes the details of the overlapping peptide pairs detection method. The method primarily consists of four procedures: *preprocessing*, *filtering*, *matching peaks and*



sequence: VTC(Cam)VVVDISKD



sequence: VTC(Cam)VVVDISK

Figure 3.2: Spectra of the peptide VTC(Cam)VVVDISKD (above) and VTC(Cam)VVVDISK (below) share the most b-ions peaks (some fragment peaks are missing during experiment)

VTC(Cam)VVVDISKD		VTC(Cam)VVVDISK	
b_1	100.07574	b_1	100.07574
b_2	201.12341	b_2	201.12341
b_3	361.15260	b_3	361.15260
b_4	460.22101	b_4	460.22101
b_5	559.28943	b_5	559.28943
b_6	658.35784	b_6	658.35784
b_7	773.38478	b_7	773.38478
b_8	886.46885	b_8	886.46885
b_9	973.50087	b_9	973.50087
b_{10}	1101.59584	-	-

Table 3.1: Theoretical b-ion masses of VTC(Cam)VVVDISKD and VTC(Cam)VVVDISK

VTC(Cam)VVVDISKD		VTC(Cam)VVVDISK		Difference
y_1	134.04483	-	-	-
y_2	262.13979	y_1	147.11285	115.02694
y_3	349.17182	y_2	234.14488	115.02694
y_4	462.25588	y_3	347.22894	115.02694
y_5	577.28283	y_4	462.25588	115.02695
y_6	676.35124	y_5	561.32430	115.02694
y_7	775.41965	y_6	660.39271	115.02694
y_8	874.48807	y_7	759.46112	115.02695
y_9	1034.51725	y_8	919.49031	115.02694
y_{10}	1135.56493	y_9	1020.53799	115.02694

Table 3.2: Theoretical y-ion masses of VTC(Cam)VVVDISKD and VTC(Cam)VVVDISK

shifted peaks detection and *scoring*.

3.2.1 Preprocessing

In proteomics samples, we observe that immonium ions occur very frequently. Immonium ions are special ions containing a single amino acid. Thus, immonium ions with the same amino acids have the same mass. If ignored, they would be recognized as matching peaks and would in turn affect the results of overlapping peptide detection. Therefore, all immonium ions are removed.

Secondly, in our algorithm for detecting overlapping peptides, we assume all peaks are charge-one peaks. Thus, all fragment peaks with two or more charges are removed.

In addition, to reduce the disturbance of noise peaks, only 15 peaks within a 100 Da window are kept and all other peaks are treated as noise peaks.

3.2.2 Filtering

Before proceeding to the detection step, two other obstacles must be addressed.

As stated above, PTMs frequently occur in proteomics. In our approach of finding overlapping peptide pairs, more matching peaks and shifted peaks between two spectra indicates higher chances of overlapping. Compared with the original peptide, the spectra of peptides with few amino acids modified would have a large number of matching peaks and shifted peaks in their spectra. In this case, it is hard to distinguish overlapping peptide from PTMs. As shown in Figure 3.3, the peptide after modification and the original peptide can be easily recognized as an overlapping peptide pair.

We also noticed that mixture spectra occur quite frequently in the sample dataset. This is because, in reality, multiple peptides are selected and fragmented concurrently resulting in a single spectrum containing fragment peaks from multiple peptides. Due to the limitation of the instrument, mass spectrometry cannot distinguish such spectra from normal spectra. Certainly, such spectra should not be considered during our selection.

To avoid these two pitfalls, we limited the mass difference of two overlapping peptides to be within a set C, which consists of all amino acid masses in addition to all combinations of two amino acid masses. We recognize two peptides as overlapping only if their sequence lengths differ by at most two. From our observation, such overlapping peptides are most common among all overlapping peptides. Since the masses of Leucine(L) and Isoleucine(I)

are the same, there are 19 different amino acid residue masses. The combination of any two amino acid residues would contain 361 different masses. In total, we obtain set C containing 380 elements.

It would be very unusual for the mass difference of two unrelated peptides to equals one element from the set C. Thus this removes the disturbance from the mixture spectra.

It should be noted that this approach does not ensure removal of all disturbances caused by PTMs. The reason is that mass differences caused by certain modifications is exactly the same as certain amino acid residue masses. For example, the Cysteine carbamidomethylation adds 57.02 Da to the original peptide. The mass change of this modification is equal to the mass of an additional Glycine. Our filtering step is not able to recognize whether two peptides differ in mass by the presence of one Glycine or by the presence of one Cysteine carbamidomethylation.

Figure 3.3: Overlapping Peptide Pair vs. PTMs

3.2.3 Matching Peaks and Shifted Peaks Detection

After all the preprocessing steps, matching peaks and shifted peaks are sought. We consider two peaks to be matched if their m/z difference is within a tolerance range of 0.03 Da. The tolerance for shifted peaks is 0.05 Da. Note that these two figures might vary for a different set of experimental data.

3.2.4 Scoring

Finally, we build a scoring function to select candidate overlapping peptides. First, all peak intensities are normalized. We set the maximum peak of a spectrum to be 1. All other peaks are linearly normalized in proportion to the maximum peak.

We consider both the quality and quantity of matching peaks and shifted peaks. We see that some of the spectra contain one or more peaks that have a much higher intensity than other peaks. Figure 3.4 shows the spectra of sequence TTPPSVYPLAPG. The peak of the y_2 ion is significantly higher than the others, which makes the score highly dependent on y_2. Therefore, we decided to smooth the intensity by computing a logarithm. Since our normalized intensity is always equal to or less than 1, to avoid having a negative value of the logarithm, after removing all values less than 0.05, we multiplied all intensity value by 200.

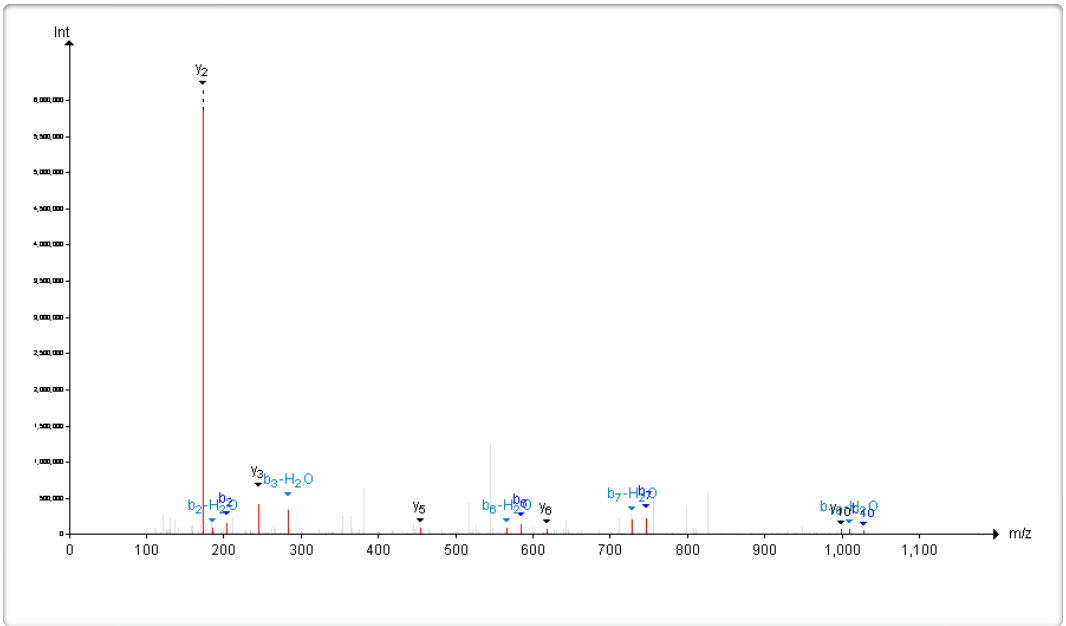


Figure 3.4: The y_2 peak significantly larger in the spectrum of sequence TTPPSVYPLAPG

For a pair of peaks, we prefer the case in which they have similar intensities rather than a major difference. Therefore, we subtract the difference of the logarithm intensity values from the lower logarithm intensity value of two peaks. Certainly, if the subtraction result is less than 0, we ignore this pair of peaks rather than adding a negative value.

We use $m_{11}, m_{12}, ..., m_{1j}$ to denote the normalized intensities of all matching peaks of the first spectrum and $m_{21}, m_{22}, ..., m_{2j}$ to indicate the normalized intensities of all matching peaks of the second spectrum. Meanwhile, $s_{11}, s_{12}, ..., s_{1k}$ represent the normalized intensities of all shifted peaks of the first spectrum, and $s_{21}, s_{22}, ..., s_{2k}$ represent the normalized intensities of all shifted peaks of the second spectrum. $m(i)$ is the score of the ith pair of matching peaks and $s(i)$ is the score of ith pair of shifted peaks. We thus arrive at the following scoring function:

$$m(i) = \min(\log(m_{1i}), \log(m_{2i})) - |\log(m_{1i}) - \log(m_{2i})|$$

$$s(i) = \min(\log(s_{1i}), \log(s_{2i})) - |\log(s_{1i}) - \log(s_{2i})|$$

$$Score = \sum_{i=1}^{j} \max(m(i), 0) + \sum_{i=1}^{k} \max(s(i), 0) \tag{3.1}$$

A preset threshold θ is used. If the score is higher than θ, we consider this pair of peptides as overlapping peptide pairs.

3.2.5 Overlapping Peptide Detection Algorithm

Based on the process described above, we build an algorithm to search for overlapping peptides.

```
   Data: spectra dataset P, threshold θ
   Result: a list L of overlapping peptide pairs
1  C ← all possible combination masses of one or two amino acid
     residue ;
2  L ← ∅ ;
3  preprocess spectrum P ;
4  for every p1 in P do
5      for every p2 in P where p1 ≠ p2 do
6          if |p1.precursor_mass − p2.precursor_mass| ∈ C then
7              M ← all matching peaks of p1 and p2 ;
8              S ← all shifted peaks of p1 and p2 ;
9              if score(M, S) > θ then
10                 Add (p1, p2) to L ;
11             end
12         end
13     end
14 end
15 return L
```

Algorithm 1: Overlapping Peptides Detection

In the algorithm, line 3 is the preprocessing procedure described in Section 3.2.1. Line 6 is the filtering procedure in Section 3.2.2. Lines 7 and 8 are the matching peak and shifted peak searches described in Section 3.2.3. Line 9 is the scoring function detailed in Section 3.2.4.

3.3 Evaluation

This section defines the test group and presents the results of overlapping peptide detection.

3.3.1 Experiment Data

The data produced by Waters that we used to evaluate the overlapping peptide detection is derived from an antibody-heavy-chain WlgG1 protein. The sample was digested by Trypsin and fragmented by the Higher-energy collisional dissociation (HCD) technique. The mass analyzer is the Fourier transform analyzer. The fragment ion error tolerance of

the instrument is 0.03 Da and the precursor ion error tolerance is 15 ppm. The spectrum data file format is *.mgf*.

3.3.2 Test Group

In total, the test dataset contains 5227 spectra. 1342 of these 5227 spectra were identified through database searching along with manual interpretation. Because identification results generated by protein database searching are reliable, we assume that these 1342 identified results are true. They are used to test the effectiveness of overlapping peptide pairs detection. Algorithm 2 was developed in order to build the test dataset. Most of the steps in Algorithm 2 are the same as in Algorithm 1. In the experiment, we take only 1342 identified spectra as the initial spectra set, as we cannot evaluate the overlapping peptides detection result if any unidentified spectra are included.

Data: spectra set E
Result: a list L of overlapping peptide sequence pairs
$C \leftarrow$ all possible combination mass of one or two amino acid residue ;
$L \leftarrow \emptyset$;
for *every* e_1 *in* E **do**
  for *every* e_2 *in* E **where** $e_1 \neq e_2$ **do**
    if $|e_1.precursor\_mass - e_2.precursor\_mass| \in C$ **then**
      if e_1 *is prefix and suffix of* e_2 **then**
        Add (e_1, e_2) to L ;
      else if e_2 *is prefix and suffix of* e_1 **then**
        Add (e_1, e_2) to L ;
    end
  end
end
return L

Algorithm 2: Overlapping Peptide Test Set Construction

3.3.3 Result

Different thresholds θ affects the results of the experiment. A higher threshold makes the program more selective. Thus, the precision of the result is increased in exchange for

reducing the total number of overlapping peptide detected. A lower threshold would cause more peptides to be recognized as overlapping and thus reduces the accuracy of detection. It is important to balance the precision and recall of the program. We have set up an experiment to determine the value of θ.

First, from the true sequences, we detected the true overlapping peptide pairs using the string matching technique. Second, we run overlapping peptide detection experiments with different θ values. Third, by comparing the detected overlapping peptide pairs with the true overlapping peptide pairs, we calculated the precision and the recall.

We restricted the precision of our method to be greater than 90 %. We found that a appropriate choice of θ is 80, at which the precision equals 91.15% and recall equals 43.53%.

In addition, to evaluate the performance of our scoring function (as shown in Function 3.1), we compared it with another scoring function (as shown in Function 3.2).

$$m(i) = \min(m_{1i}, m_{2i}) - |m_{1i} - m_{2i}|$$

$$s(i) = \min(s_{1i}, s_{2i}) - |s_{1i} - s_{2i}|$$

$$Score = \sum_{i=1}^{j} \max(m(i), 0) + \sum_{i=1}^{k} \max(s(i), 0) \tag{3.2}$$

Function 3.2 is similar to Function 3.1 except that it does not take the logarithm of the intensity of peaks.

Finally, we plotted precision-recall curves. As shown in the result, the scoring function with the logarithm slightly outperforms the one without the logarithm, which implies that smoothing the intensities of the peaks is helpful.

For a desktop computer with 16 GB of memory, the overlapping peptides detection program takes approximately 2 seconds to run. As a comparison, Novor takes approximately 17 seconds to identifying 1342 spectra using the same computer. Since overlapping

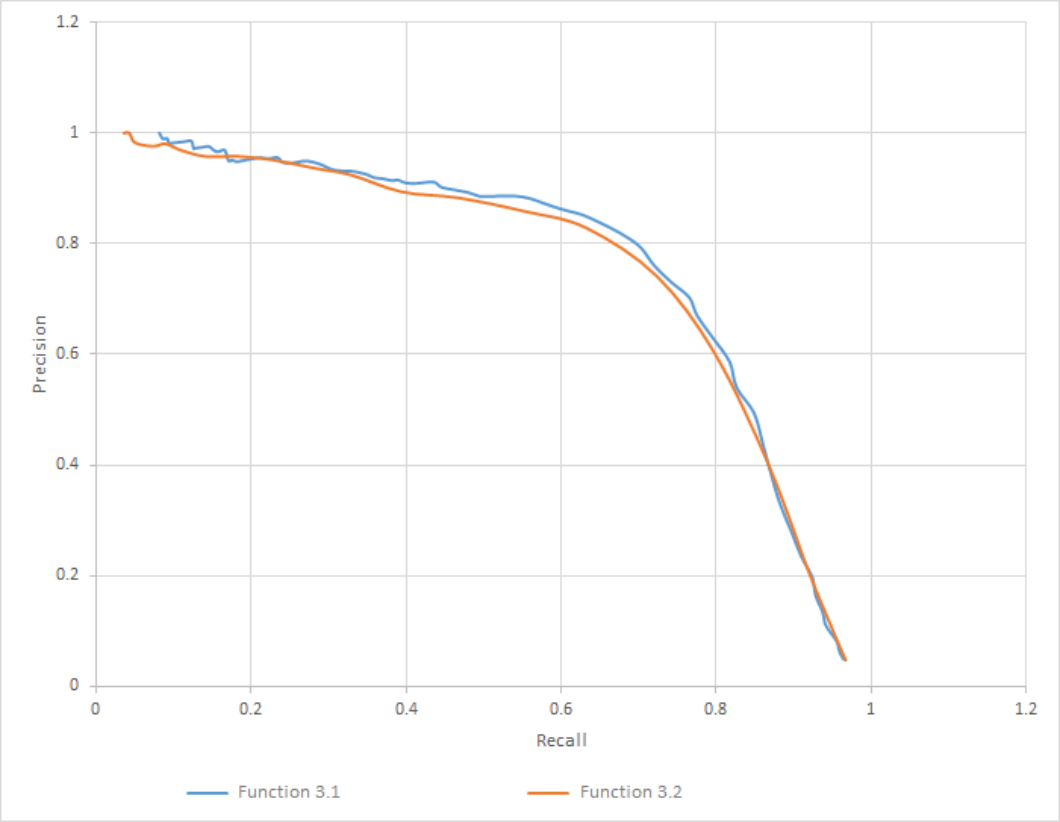


Figure 3.5: The precision-recall curve of the overlapping peptide pairs detection for different scoring functions

peptides detection can be run in parallel with Novor, we conclude that overlapping peptides detection should not affect the total running time of Novor for a spectrum dataset of similar size.

Chapter 4

Labeling

4.1 Misclassified Peaks in De Novo Sequencing

One of the difficulties in de novo sequencing is that a contiguous ion series might be identified, but the direction of the sequence may be difficult to establish. In other words, for CID data it may not be clear whether an ion series is y-type or b-type [29]. We therefore devised an experiment to test how many peaks are misclassified in an incorrect identification result.

The test dataset we used is the same as that described in Section 3.3. We used all 1342 identified peptides for this experiment and we used Novor for de novo sequencing. For each spectrum, we have a de novo sequencing result and a database searching result. If the two results match, we considered the de novo sequencing result as a correct identification, and if not, we considered it to be an incorrect identification. To test for misclassified peaks, we calculated all theoretical b-ion peaks from database searching result sequences and calculated all theoretical y-ion peaks from de novo result sequences. We then check these two groups of peaks and observed whether there is a match. There are two conditions that can cause a match in this experiment. The first one is that Novor misclassifies a fragment ion peak. The second one is that a b-ion peak coincidentally has exactly the same mass as a y-ion peak. We divided the test group into a correct identification group and an incorrect identification group and counted the total number of matches for each group. The results are presented in Table 4.1.

For the incorrect identification group, clearly, Novor does not misclassify any ion peaks. The only reason for a match is that a b-ion peak coincidentally has exactly the same mass

	# of spectrum	# of match in total	rate
correct	308	14	0.045
incorrect	1034	659	0.637

Table 4.1: Comparison of the number of misclassified peaks between the correct identification group and the incorrect identification group

as a y-ion peak. The occurrence rate of this situation is around 0.045 match/spectra. In the incorrect identification group, the probability of coincidences should be similar to the probability for the correct identification group. From the experiment, however, we observed a large difference in the match rate between the two groups (0.592 match/spectra). We conclude that the incorrect direction of peaks classification occurs frequently in incorrect de novo results.

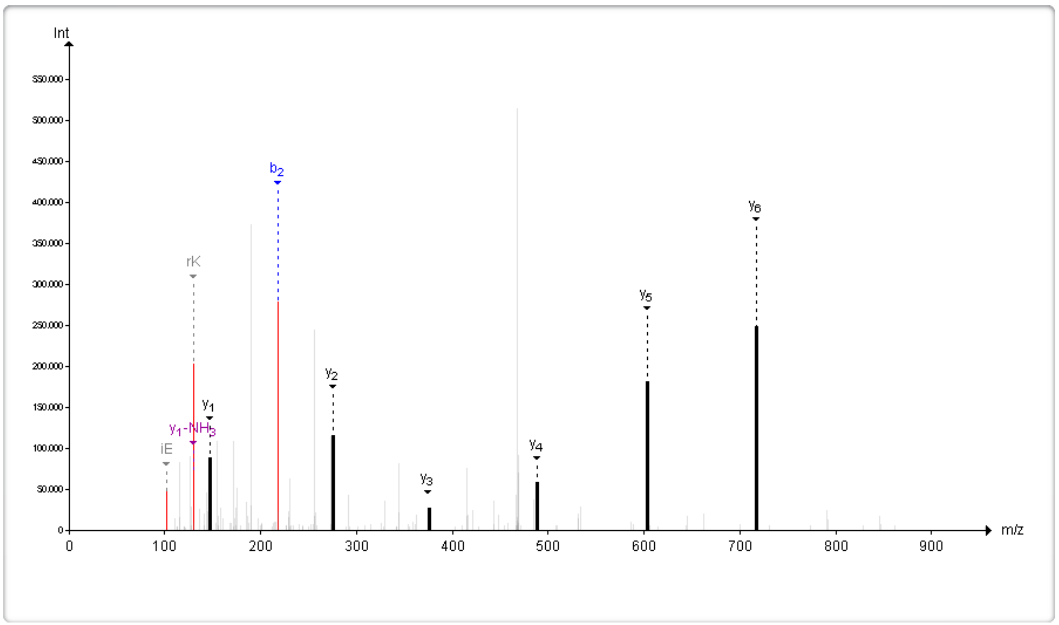
Therefore, reducing the number of misclassified peaks would greatly improve the de novo result. Using overlapping peptides, we are able to achieve this goal.

For misclassified spectra, a single spectrum does not usually contain sufficient information to directly identify b-ion peaks or y-ion peaks. However, overlapping peptide pairs would provide more information. As shown in Figure 3.2, sequences VTC(Cam)VVVDISKD and VTC(Cam)VVVDISK are overlapping. Matching peaks contain the most b-ion peaks while shifted peaks contain the most y-ion peaks. Figure 4.1 shows another example, in which spectra from sequences SEIDNVKK and LRSEIDNVKK share the most y-ion peaks(highlighted by black). Matching peaks contain the most y-ion peaks and shifted peaks contain the most b-ion peaks.

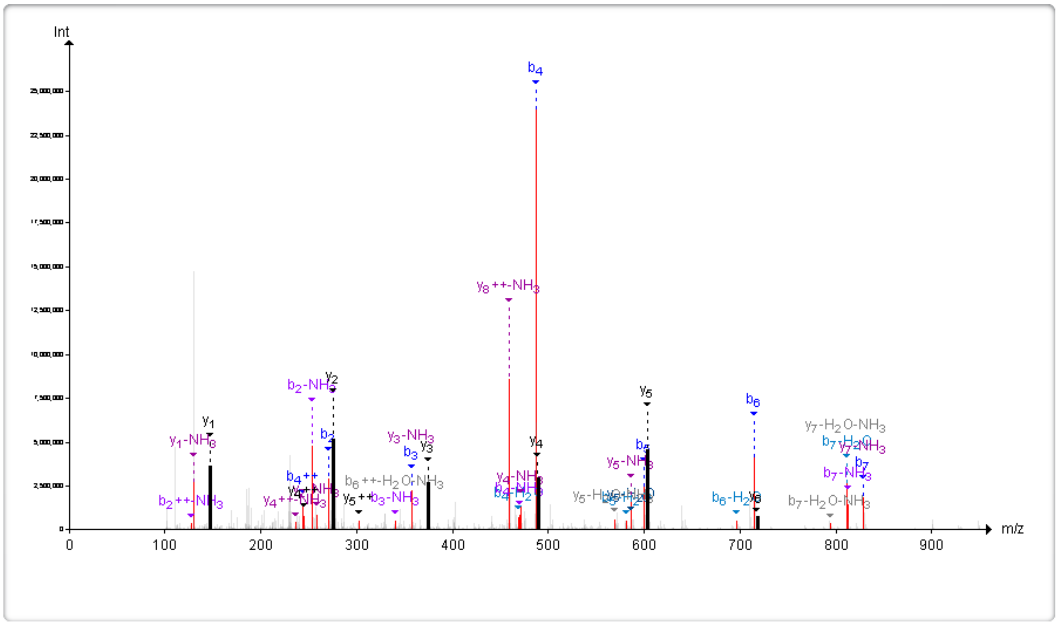
In other words, if we have an overlapping peptide pair, we can separate the fragment peaks into three groups: matching peaks, shifted peaks and others. Most b-ion peaks and y-ion peaks could then be separated accordingly. Although we cannot directly tell which group are the b-ion peak group, such separation will certainly assist de novo sequencing.

4.2 Methodology

This section introduces the method of correcting de novo sequencing errors by reducing misclassified peaks. The method consists of two main parts: spectrum labeling and Novor modification. In the spectrum labeling part, the peaks of spectra are labeled and a new spectrum data file is created. In the Novor modification part, the Novor source code is modified to adopt the new labeled spectrum data file.



seq: SEIDNVKK



seq: LRSEIDNVKK

Figure 4.1: Spectra of peptide SEIDNVKK (above) and LRSEIDNVKK (below) share the most y-ions peaks (some fragment peaks are missing during the experiment)

4.2.1 Spectrum Labeling

The main purpose of this part is to generate a new spectrum file with labeled peaks. As described in Section 4.1, most b-ions and y-ions can be separated by searching for matching peaks and shifted peaks. Therefore, by labeling matching peaks and shifted peaks, b-ion peaks and y-ion peaks can be labeled in different peak groups.

First, we modified Algorithm 1 to find only the most possible overlapping peptide for each spectrum. We refer to this peptide as its *spouse*. Note that, unlike the literal meaning of spouse, the spouse in this thesis is not a reflexive relationship. In other words, if spectrum A's spouse is spectrum B, spectrum B's spouse is not necessarily spectrum A. By modifying Algorithm 1, we restrict each spectrum to have only zero or one spouse.

Data: spectra dataset P, threshold θ
Result: an array A of spouse information for each spectrum
$C \leftarrow$ all possible combination mass of one or two amino acid residue ;
$A \leftarrow$ an empty array with length equals the number of P ;
preprocess spectrum P ;
for *every* p_1 *in* P **do**
  for *every* p_2 *in* P **where** $p_1 \neq p_2$ **do**
    if $|p_1.precursor\_mass - p_2.precursor\_mass| \in C$ **then**
      $M \leftarrow$ all matching peaks of p_1 and p_2 ;
      $S \leftarrow$ all shifted peaks of p_1 and p_2 ;
      $score \leftarrow$ score(M, S) ;
      if $score > \theta$ **then**
        if $A[p_1]$ *is not initialized* **or** $A[p_1].score < score$ **then**
          $A[p_1].score \leftarrow score$;
          $A[p_1].spouse \leftarrow p_2$;
        end
      end
    end
  end
end
return A

Algorithm 3: Best Spouse Searching

Algorithm 3 returns an array rather than a list. The overlapping peptide with the highest score to be the spouse of spectrum p_1 is selected in steps 11 to 13.

The labeling procedure is then relatively simple, as shown in Algorithm 4.

Data: spectra dataset P, array A of spouse information for each spectrum produced by Algorithm 3
Result: a spectrum file F of labeled spectrum

```
1  F ← empty file ;
2  for every p1 in P do
3      if A[p1] is initialized then
4          p2 ← A[p1].spouse ;
5          M ← all matching peaks of p1 and p2 ;
6          S ← all shifted peaks of p1 and p2 ;
7          Remove common peaks from M and S ;
8          Label all peaks of M as 'M' ;
9          Label all peaks of S as 'S' ;
10         Append labeled spectrum p1 to F ;
11     end
12 end
13 return F
```

Algorithm 4: Spectrum Labeling

In Algorithm 4, all matching peaks and shifted peaks are marked except for peaks appearing in both groups (step 7). Peaks appearing in both groups cannot be treated only as matching peaks or only as shifted peaks, so they are not labeled.

A new file format *.mgfl* has been created on the basis of the *.mgf* format. The only difference from *.mgf* is that a label "M" or "S" is added to the ion list. Figure 4.2 shows an example of an *.mgfl* file.

4.2.2 Novor Modification

We modified the Novor software to handle the labeled spectra. After the spectrum labeling, all peaks are divided into three groups: unlabeled peaks, peaks marked as "M" and peaks marked as "S".

In reality, some fragment ion peaks might not be presented in the spectra. As a result, for certain fragment ion peaks, corresponding matching or shifted peaks might be missing,

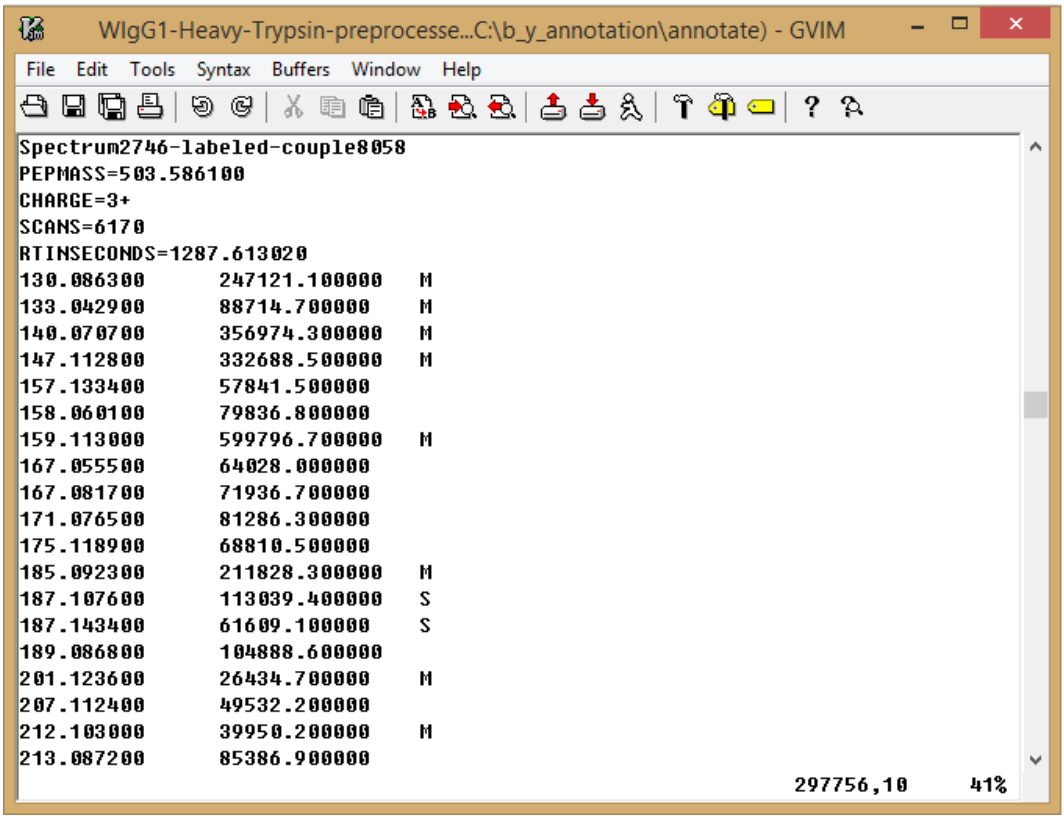


Figure 4.2: An example of the labeled spectrum file

which causes those fragment peaks to be unlabeled. Therefore, for unlabeled peaks, we cannot ignore the possibility that they are fragment peaks rather than noise.

Novor is modified to treat a cluster of labeled peaks only as b-ion peaks or y-ion peaks. For example, for all peaks labeled as "M", Novor considers them only as b-ion peaks which include b-ion fragment peaks and their $-NH_3$ peaks, $-H_2O$ peaks etc. Novor treats all unlabeled peaks as it normally does since all unlabeled peaks may be either b-ion peaks or y-ion peaks.

For each spectrum, Novor is then run three times. First, the labeling feature is ignored and Novor is run as usual. Second, all peaks labeled "M" are treated as b-ion peaks and all peaks labeled "S" are treated as y-ion peaks. Third, all peaks labeled "S" are treated as b-ion peaks and all peaks labeled as "M" are treated as y-ion peaks. The sequencing result with the highest confidence score calculated by Novor is selected. In addition, we

export all three sequencing results and their confidence scores for further investigation.

4.3 Evaluation

4.3.1 Experiment Data

We use the same experiment dataset as in Section 3.3.

4.3.2 Test Group

Again, we use only 1342 out of 5227 identified spectra to examine our result. We consider a match between the de novo sequencing result and the database searching result to be a correct sequencing.

In theory, by directly comparing the sequencing result before and after the labeling, we will see the effect of the method. However, at the time the experiment was conducted, Novor only supported a few types of PTMs. For some peptides with PTMs not covered by Novor, Novor produces inaccurate results. For example, Novor does not support Carbamidomethyl on the N-terminus. A sequence such as (N-term|Cam)GQPAENYK would not be correctly identified. Instead, Novor will produce GGQPAENYK since the modification of Carbamidomethy on the N-terminus would add 57.02 Da to the N-terminus, which is similar to adding a Glycine to the front of the sequence.

For this reason, we developed another way to test our results. We mapped a sequence to a residue mass array. To make the description more intuitive, we use a segmented line to represent the mass array. As shown in Figure 4.3, each point represents the sum of residue masses of the previous amino acids. We mapped results generated by the database searching and results from de novo sequencing to arrays. Then, for each database searching result sequence, we checked every amino acid to determine whether its two boundaries appear in the corresponding de novo sequencing result array. Finally, we counted the total number of matched amino acids. In the example shown in Figure 4.3, every amino acid of (N-term|Cam)YYC(Cam)TR has its boundaries appear in the array of sequence GYYC(Cam)TR even though Carbamidomethyl on the N-terminus is not correctly interpreted.

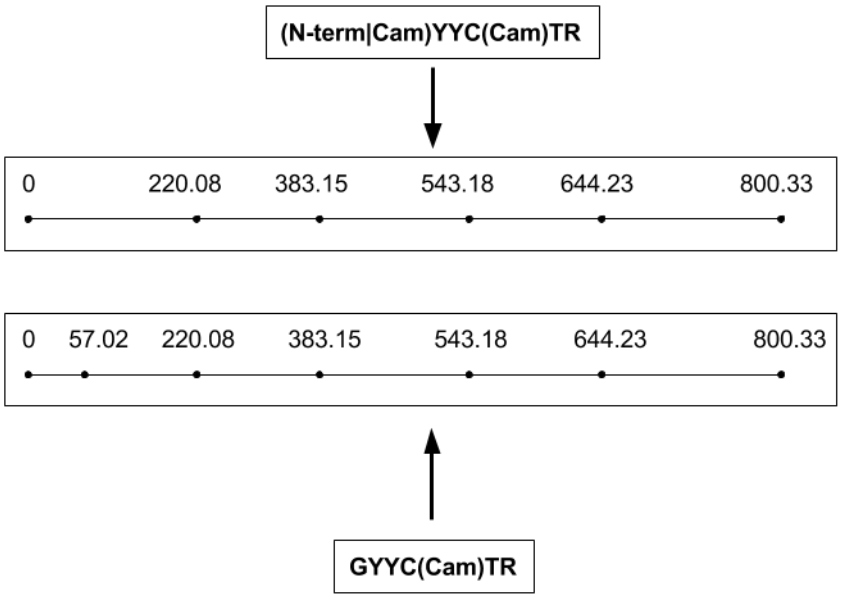


Figure 4.3: Mapping a sequence into an array of residue masses

4.3.3 Result

By comparing the number of matched amino acids, we can evaluate the performance of the labeling method. In the experiment, we set the threshold θ to be 80. In Figure 3.5, the threshold of 80 is the point at which the recall equals 43.53% and the precision equals 91.15%. We favor precision over recall. This experiment includes 650 identified peptides that are labeled. There are 11521 amino acids in total. For the result produced by the modified version of Novor, we count the number of correctly interpreted amino acids. The result is shown in the first column of Table 4.2.

In addition, we set up another experiment to test the best potential improvement. Instead of selecting the sequence with the most confidence scores, we selected the best sequencing results that have the most matched amino acids compared with the true peptide sequences. Since we consider database searching results as true peptide sequences, we pick the one with the most matched amino acids based on the database searching result. The

	Sequence with highest score	Sequence with best matches
Total labeled peptide	650	650
Total AA	11521	11521
Matched AA before the labeling	7033	7033
Matched AA after the labeling	7092	7327
Improvement Ratio(%)	0.84	4.18
Peptides become better	47	117
Peptides become worse	32	0

Table 4.2: Result of labeling under current confidence scoring function and theoretical theoretical improvement of a different scoring function

result is shown in the second column of Table 4.2.

Novor runs three times for labeled spectra. Despite the long running time, the actual result improvement is not remarkable being only 0.84%. A possible reason is that the current Novor confidence score function does not take the labeling feature into consideration. As shown in Table 4.2, we are able to reach 4.18% potential improvement. Another reason is that fragment ion peaks classification has already been taken into consideration in Novor. The labeling strategy might duplicate the effort of Novor.

Chapter 5

Voting

5.1 Correcting De Novo Results with Low aaScore

As discussed in [29], de novo sequencing often produces partially correct sequences. A strategy to increase the accuracy of peptide identification is to combine de novo sequencing with database searching. In this chapter, we propose another strategy to deal with partially correct sequences.

In addition to a peptide confidence score, Novor also produces an Amino Acid Confidence Score (aaScore) for each identified spectrum. The aaScores list is a list of scores representing the confidence of each amino acid. They range from 0 to 99. Larger scores stand for more confidence that this amino acid is correctly interpreted. For example, Novor correctly identified the sequence GQPAENYK and its aaScores are 78-91-99-96-96-95-82-70 (first row in Table 5.1). All aaScores in the sequence are high. However, for some other spectra, Novor produces results with lower aaScores such as the sequence GGQPAENYK (second row in Table 5.1). The sequence GQGPAENYK is a partially correct result produced by Novor. By comparison with the true sequence, the second and the third amino acid are incorrectly interpreted while all others are accurate. In addition, the aaScores of the second and the third amino acid are low. Such partially correct results are typically caused by low-quality spectra such as those with fragment peaks missing.

Novor currently processes each spectrum independently. In this thesis, we have constructed relations between spectra using overlapping peptides. By utilizing overlapping peptides, we are able to supply additional information to Novor and thus improve the accuracy.

True Sequence	Novor Result Sequence	aaScore
GQPAENYK	GQPAENYK	78-91-99-96-96-95-82-70
GGQPAENYK	GQGPAENYK	85-12-3-91-98-96-97-91-92

Table 5.1: Novor's interpretation of an overlapping peptide pair

We realized that two sequences in Table 5.1 are considered to overlap. The first sequence is the suffix of the second one. By aligning two de novo result sequences and taking the aaScores into consideration, we are able to modify the partially correct result by reversing the order of the second and the third amino acid.

The main method of detecting overlapping peptides is described in Chapter 3. In this chapter, we show the method for replacing incorrect subsequences in Novor results.

5.2 Methodology

Voting consists of two steps: alignment and replacement. In the alignment step, it is determined whether the shorter sequence is the prefix or suffix of the longer sequence. In addition, two peptides are aligned. In the replacement step, amino acids with lower aaScores are replaced according to their alignment.

5.2.1 Alignment

Similar to the spectrum labeling method introduced in Section 4.2.1, we restrict each spectrum to only zero or one spouse.

Two significant problems arise when given an overlapping peptide pair. First of all, it is important to decide whether the two sequences share the same prefix or share the same suffix. A wrong decision would certainly make all the ensuing procedures invalid. The second problem is deciding how amino acids are replaced. A basic rule is that we need to keep the residue mass of the sequence unchanged after replacement. Clearly, we cannot replace an Alanine(A) with an Arginine(R) since residue masses of A and R are different. However, we could replace two Glycine(G) with an Asparagine(N) since their residue masses are be same. We could replace [AQG] by [GAQ] since the latter is a permutation of the former.

To resolve these two problems, we make an alignment of the overlapping peptide pair. Since the shorter sequence is either the prefix or suffix of the longer sequence, we try both

cases and select the better one. We add a mass gap to either the head or tail of the shorter sequence in order to match the total residue mass of the longer sequence. Then, similar to the method of mapping a sequence to an array as described in Section 4.3.2, we map two sequences into two arrays of residue mass. By aligning the two arrays of residue masses, alignment with more overlapping masses indicates a higher probability of being the correct one. Finally, we count the number of overlapping masses in the two arrays and select the larger one.

After alignment, the sequences are partitioned into a number of segments. Each segment from one peptide has the same residue mass as the corresponding segment from the other peptide. These segments are used in the next step for replacing amino acids.

The overall flow of alignment is shown in Figure 5.1.

The dark circle in Figure 5.1 represents a gap. The length of the gap is equal to the precursor mass difference of the two sequences. We insert the gap to either the head or the tail of the shorter sequence. When the gap is inserted to the head, there are 10 overlapping masses, which is larger than the number of overlapping masses when the gap is inserted to the tail. Thus, we choose the left alignment and the sequences are partitioned into 9 segments.

The detail procedure of aligning two residue mass arrays is shown in Algorithm 5.

```
   Data: mass array l_1 and l_2
   Result: an alignment t of l_1 and l_2
 1 i_1 ← 0 ;
 2 i_2 ← 0 ;
 3 t ← ∅ ;
 4 while i_1 < l_1.length or i_2 < l_2.length do
 5 |   if l_1[i_1] overlaps l_2[i_2] then
 6 |   |   Add index i_1 and i_2 to t ;
 7 |   |   Increment i_1 ;
 8 |   |   Increment i_2 ;
 9 |   else if l_1[i_1] > l_2[i_2] then
10 |   |   Increment i_2 ;
11 |   else
12 |   |   Increment i_1 ;
13 end
14 return t
```

Algorithm 5: Alignment

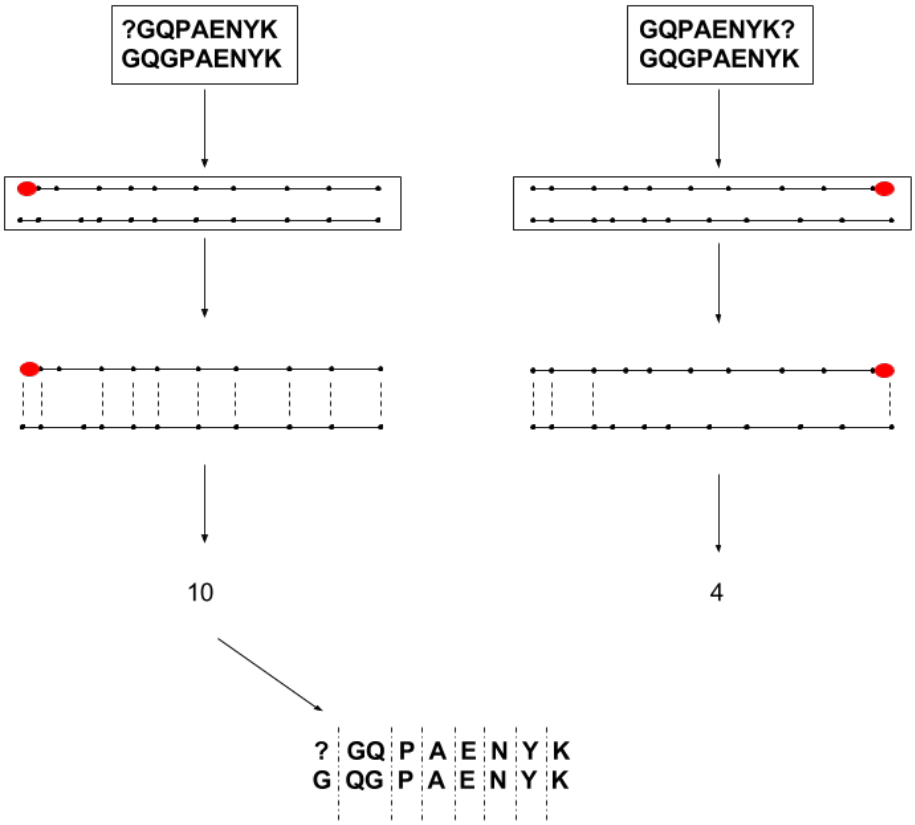


Figure 5.1: Flow of alignment

5.2.2 Replacement

After generating an alignment of sequences, the sequences are also partitioned into a number of segments. The replacement step is based on segments.

In the sequence GQGPAENYK in Table 5.1, the aaScores of the second and the third amino acid are low. This is shown in Figure 5.2. After aligning with the sequence ?GQPAENYK where the mass of "?" is 57.02, we replace the second segment QG in GQGPAENYK with the second segment GQ in ?GQPAENYK. Since the masses of corresponding segments from two sequences are same, we are able to safely replace them. So far, we have successfully corrected the second and the third amino acids in the sequence GQGPAENYK. The remaining two sequences are exactly the same, so we leave them unchanged.

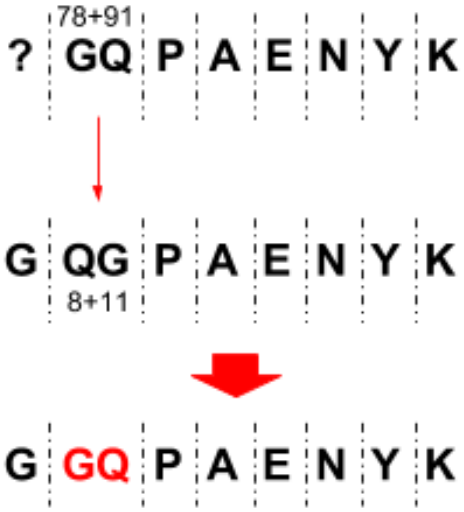


Figure 5.2: Flow of replacement

```
   Data: alignment t, multiplier α
 1 for every pair of segments (seg_1, seg_2) in t do
 2 |   if seg_1 not equal seg_2 then
 3 |   |   if seg_1 not contains gap and seg_2 not contains gap then
 4 |   |   |   x_1 ← total aaScore of seg_1 ;
 5 |   |   |   x_2 ← total aaScore of seg_2 ;
 6 |   |   |   if x_1 > α · x_2 then
 7 |   |   |   |   Replace seg_2 by seg_1
 8 |   |   |   end
 9 |   |   |   if x_2 > α · x_1 then
10 |   |   |   |   Replace seg_1 by seg_2
11 |   |   |   end
12 |   |   end
13 |   end
14 end
```

Algorithm 6: Replacement

Algorithm 6 presents the detailed explanation of how segments are replaced. Alignment structure t contains all the information needed for replacement such as the identified sequences, the aaScores of each amino acids and the segment array of alignments. We also set a multiplier α. Whenever the total aaScores of one segment is α times greater than the other, we substitute the lower one. In Step 3, we ensure that none of segments contains a gap. If there is a gap, it is meaningless to substitute segments. Steps 7 and 10 are the replacement procedures.

5.3 Evaluation

5.3.1 Experiment Data

In addition to the sample data we used in Section 3.3.1, we include more samples in order to evaluate the voting strategy. We use antibody samples RN-R206-IS and RN-R207-IS. These samples are digested by AspN, Chymotrypsin, Pepsin and Trypsin respectively, forming eight spectrum datasets in total. These eight groups of data have been produced recently and thus more convincing in terms of the evaluation strategy.

5.3.2 Test Group

We use the same method in Section 4.3.2 to test the improvement of Novor after applying the voting strategy. We treat the result sequences generated by the protein database searching technique as the true sequences. The numbers of matched amino acids between the de novo sequencing results and database searching results are the main indicator of measurement. For all sample data, we select the spectra that were identified through database searching as input for the voting program.

5.3.3 Result

Different choices of multiplier α in Algorithm 6 and threshold θ in Algorithm 3 can affect the results of the experiment. A greater value of α indicates more caution when the algorithm decides whether to modify the result sequences and thus leads to fewer sequences being corrected and fewer mistakes being made. Threshold θ, described in Section 3.2.4, affects the total number of overlapping peptides detected.

Therefore, we set up an experiment to test the effects from different values for θ and α. We again use the 1342 identified spectra from WlgG1 data and calculate the improvement ratio as follows:

$$improvement\ ratio = \frac{matched\ AA\ after\ the\ voting\ - matched\ AA\ before\ the\ voting}{matched\ AA\ before\ the\ voting}$$

Table 5.2 shows the results for different values of α with θ at 80. In the overlapping peptide detection experiment, when θ equals 80 the precision equals 0.9115 and recall equals 0.4353.

From Table 5.2, we see that as α increases, the number of modified peptides decreases. However, the improvement decreases. The improvement ratio reaches its maximum when α equals 1, which implies that the algorithm should act more aggressively. When a pair of different segments is found, the one with the lower aaScore should always be replaced.

Table 5.3 shows the different values of θ with α set at 1. As seen in Table 5.2, the algorithm performs best when α equals 1. Since different values for θ restrict the total number of overlapping peptides detected, we define the affected peptide ratio. The term *affected peptides* denotes peptides with spectra that match with a "spouse". Only spectra with a "spouse" are modified in the voting. The affected peptide ratio is calculated as follows:

$$affected\ peptide\ ratio = \frac{affected\ peptides}{total\ number\ of\ peptides}$$

Choice of α	1	3	6	10
Total labeled peptide	639			
Total AA	11374			
Matched AA before the voting	6862			
Matched AA after the voting	7207	7034	6956	6928
Improvement Ratio (%)	5.03	2.51	1.37	0.96
Peptide become better	106	42	26	20
Peptide become worse	32	2	1	0

Table 5.2: Result of Voting when α equals 1, 3, 6 and 10

Choice of θ	40	80	120	160
Precision of experiment in Section 3.3	0.6275	0.9115	0.9509	0.9897
Recall of experiment in Section 3.3	0.7968	0.4353	0.1915	0.0863
Total affected peptide	1099	639	400	230
Total affected AA	17451	11374	7796	4904
Affected peptide ratio (%)	81.9	47.6	29.8	17.1
Matched AA before the voting	10568	6862	4690	2936
Matched AA after the voting	10553	7207	4979	3137
Improvement Ratio (%)	-0.14	5.03	6.16	6.85
Peptide become better	95	106	81	56
Peptide become worse	86	32	22	13

Table 5.3: Result of Voting when θ equals 40, 80, 120 and 160

In Table 5.3, as θ increases, the precision of overlapping peptides detection also increases. As a result, the improvement ratio increases as well. However, the number of affected peptides decreases. We set θ at 80 for the following experiments as it balances the improvement ratio and the number of affected peptides.

After choosing the value of θ and α, we evaluated the voting method with antibody samples RN-R206-IS and RN-R207-IS. Table 5.4 shows the result.

In summary, by testing different sample data, we reached improvement ratios ranging from 3% to 8% for overlapping peptides. The voting spends 165 milliseconds correcting

Data index	Enzymes	Improvement ratio (%)	Affected peptide ratio (%)
R206	AspN	6.81	54.91
R206	Chymotrypsin	4.67	66.44
R206	Pepsin	4.79	70.88
R206	Trypsin	3.10	60.26
R207	AspN	7.27	43.38
R207	Chymotrypsin	3.59	67.96
R207	Pepsin	6.13	70.11
R207	Trypsin	3.75	62.05

Table 5.4: Experiment results for Samples R206 and R207 with $\alpha = 1$ and $\theta = 80$

the de novo sequencing result for a 1342-spectrum sample on a desktop computer with 16 GB of memory. As mentioned in Section 3.3.3, the overlapping peptide detection program takes approximately 2 seconds to run and Novor takes 17 seconds for the same sample. Since overlapping peptide detection and Novor can run simultaneously, the added time of the voting is extremely small compared with the significant improvement of the de novo sequencing accuracy.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Due to the improvement of computing algorithms and desktop computer hardware, de novo sequencing has now become practical. With the establishment of Novor, the speed of de novo sequencing can greatly outperform the speed of the protein database searching. Increasing the accuracy of de novo sequencing has become an urgent goal.

In this thesis, we focus on improving de novo sequencing using overlapping peptides. We propose a method of detecting overlapping peptides directly from the spectrum without the need for database searching.

The labeling and the voting are the two strategies we designed to improve the result from de novo sequencing using overlapping peptides.

The labeling strategy relies on separating fragment peaks to reduce the number of misclassified fragment ions. Although the evaluation experiment indicates that the labeling strategy does not produce remarkable outcomes, we still find that potential improvement can be achieved.

The voting strategy depends on substituting substrings in the Novor results. With the great improvement of accuracy and the short running time, the voting strategy was proven to be a worthy supplement of de novo sequencing. Although all experiments involving de novo sequencing were performed by Novor, the voting should be added not only to Novor but also to all de novo sequencing tools that output amino acid confidence scores.

Furthermore, the idea of voting can be applied to the method of protein database searching and other sequence identification tools.

6.2 Proposed Future Work

This section proposes topics for future studies.

6.2.1 Substitution Algorithm of the Voting

This thesis uses a straightforward substitution algorithm to replace the lower score substring. A more accurate and complicated algorithm might be developed to further improve the accuracy of de novo sequencing. For example, machine learning techniques could be added to detect an incorrect substring.

6.2.2 Overlapping Peptide Cluster

Ideally, it would be more intuitive if we could construct overlapping peptide clusters after finding overlapping peptide pairs. An overlapping peptide cluster is a group of overlapping peptides all sharing the same prefix or suffix.

To make the description more clear, we use a graph to represent spectrum data. Initially, the graph consists of a number of vertices without any edges. Each vertex represents a spectrum. By finding overlapping peptide pairs, vertices would be connected accordingly. An edge exists only if its connected vertices (spectra) overlap. In graph theory, a clique is defined as the maximal complete subgraph of an undirected graph[35]. A clique would be a perfect representative of a cluster overlapping peptides. Therefore, once all overlapping peptide pairs have been found, finding the overlapping peptide clusters is the same as finding cliques (see Figure 6.1).

There exist many obstacles to overlapping peptide cluster construction:

- Clique detection is NP-hard and thus an appropriate approximate algorithm is needed.
- In practice, some overlapping peptide pairs may not be successfully detected due to the low quality of spectra. Undetected overlapping peptides result in missing edges, which complicates the detection of the overlapping peptide clusters.

Once an overlapping peptide cluster is available, the ideas of labeling and voting can be applied to improve de novo sequencing. Instead of a single spectrum, a group of spectra must be considered together. The problem of assignment of matching and shifted peaks from different overlapping peptides has to be solved.

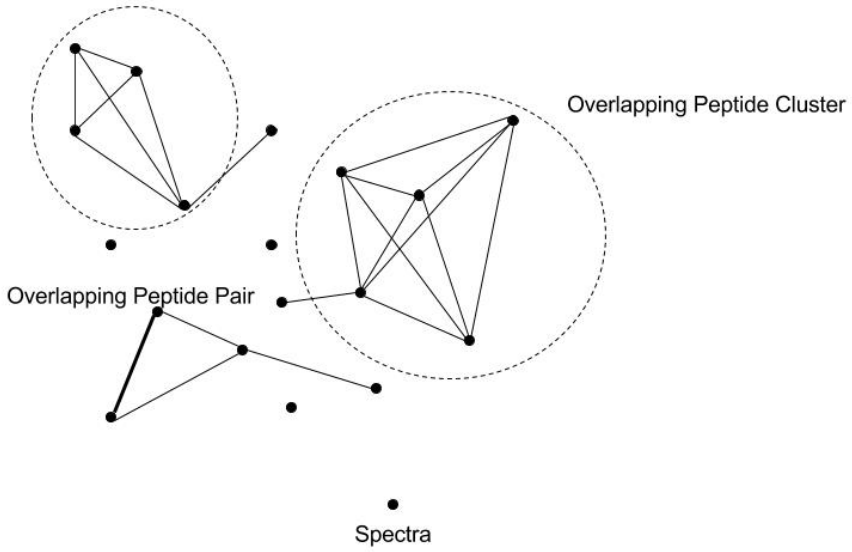


Figure 6.1: Graph representation of overlapping peptide clusters

References

[1] Ma, B. (2015) Novor: Real-Time Peptide de Novo Sequencing Software. *J. Am. Soc. Mass Spectrom.*, **26**, 1885-1894.

[2] ThermoFisher Scientific. (2017) Overview of Post-Translational Modifications. https://www.thermofisher.com/ca/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-post-translational-modification.html

[3] Rockland. (2017) Post-Translational Modification Antibodies http://www.rockland-inc.com/post-translational-modification-antibodies.aspx

[4] Sigma Aldrich. (2017) Peptide Modifications: N-Terminal, Internal, and C-Terminal http://www.sigmaaldrich.com/technical-documents/articles/biology/peptide-modifications-n-terminal-internal-and-c-terminal.html

[5] ThermoFisher Scientific. (2017) Overview of Mass Spectrometry for Protein Analysis. https://www.thermofisher.com/ca/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-mass-spectrometry.html

[6] Bioinformatics. (2017) Flow of ToF. https://www.bioinformatics.ca

[7] Dempster, A. J. (1918) A new Method of Positive Ray Analysis. *Physical Review.*, **11**, 316-325.

[8] Fenn, J.B. et al. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science.*, **246**, 64-71.

[9] Karas, M. et al. (1987) Matrix-assisted ul-traviolet laser desorption of non-volatile compounds. *Int. J. Mass Spectrom. Ion Process.*, **78**, 53-68.

[10] W. Paul, H. Steinwedel. (1953) A New Mass Spectrometer without a Magnetic Field *Z. Naturforsch.*, **8a**, 448450

[11] H.Dawson, Peter. (1976) Quadrupole Mass Spectromety and Its Application. *New York: Elsevier Scientific*

[12] Stephens, W. E. (1946) A Pulsed Mass Spectrometer with Time Dispersion *Phys. Rev.*, **69**, 691

[13] Amster, I.J. (1996) Fourier transform mass spectrometry *Journal of mass spectrometry.*, **31**, 1325-1337

[14] Zubarev, Roman A. and Makarov, Alexander. (2013) Orbitrap Mass Spectrometry *Analytical Chemistry.*, **85**, 5288-5296

[15] Matrix Science. (2017) Peptide fragmentation. http://www.matrixscience.com

[16] Wells, J. Mitchell. (2005) CollisionInduced Dissociation (CID) of Peptides and Proteins *Methods in Enzymology.*, **402**, 148-185

[17] Bioinfomatics Solution Inc. (2017) De novo Peptide Sequencing http://www.bioinfor.com/de-novo-sequencing/

[18] Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.*, **20**, 35513567

[19] Eng, J., McCormack, A. L., and Yates, J. R., 3rd (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976 989

[20] Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 13671372

[21] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.*, **3**, 958 964

[22] Craig, R., and Beavis, R. C. (2004) TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics.*, **20**, 1466 1467

[23] Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquiredon a quadrupole selecting quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in protein prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol. Cell. Proteomics.*, **4**, 1194 1204

[24] Kim, S., Mischerikow, N., Bandeira, N., Navarro, J. D., Wich, L., Mohammed, S., Heck, A. J., and Pevzner, P. A. (2010) The generating function of CID, ETD and CID/ETD pairs of tandem mass spectra: Applications to database search. *Mol. Cell. Proteomics.*, **9**, 2840 2852

[25] Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A., Lajoie G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **17**, 23372342

[26] Frank A., Pevzner P. (2005) PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, **77**, 964973

[27] Taylor J. A., Johnson R. S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **11**, 10671075

[28] Fischer B., Roth V., Roos F., Grossmann J., Baginsky S., Widmayer P., Gruissem W., Buhmann J. M. (2005) NovoHMM: A hidden Markov model for de novo peptide sequencing. *Anal. Chem.*, **77**, 72657273

[29] Ma B, Johnson R. (2012) De novo sequencing and homology searching. *Mol. Cell. Proteom.*, **11** , O111.014902.

[30] Vlado Dancik, Theresa A. Addona, Karl R. Clauser, James E. Vath, and Pavel A. Pevzner. (1999) De Novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology.*, **6**, 327-342

[31] Zhang J, Xin L, Shan B, et al. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Molecular & Cellular Proteomics: MCP.*, **11(4)**, M111.010587

[32] Wilhelm, Thomas and Jones, Alexandra M. E. (2014) Identification of Related Peptides through the Analysis of Fragment Ion Mass Shifts. *Journal of Proteome Research.*, **13**, 4002-4011

[33] Bandeira, Nuno and Tsur, Dekel and Frank, Ari and Pevzner, Pavel A. (2007) Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences.*, **104**, 6140-6145

[34] Rapid Novor. (2017) Antibody Protein De Novo Sequencing. https://www.rapidnovor.com/

[35] Luce, R. D. and A. Perry (1949) A method of matrix analysis of group structure. *Psychometrika.*, **14**, 94-116.