De novo PEPTIDE SEQUENCING METHODS FOR TANDEM MASS SPECTRA

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Division of Biomedical Engineering
University of Saskatchewan
Saskatoon

By
Yan Yan

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Division of Biomedical Engineering
Engineering Building
57 Campus Dr.
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5A9

Abstract

De novo peptide sequencing from MS/MS spectra has become of primary importance in proteomics. It provides essential information for studies of protein structure and function. With the availability of various MS/MS spectra, a lot of computational methods have been developed to infer peptide sequences from them. However, current *de novo* peptide sequencing methods still have limitations. Some major ones include a lack of suitable models reflecting MS/MS spectra, limited information extracted from MS/MS spectra, and the inefficient use of multiple spectra. This thesis addresses some of the limitations with a series of novel computational methods designed for various MS/MS spectra and their combinations.

The main content of the thesis starts with a comprehensive review of recent developments in *de novo* peptide sequencing methods, followed by two novel methods for single spectrum sequencing problems, and then presents two paired spectra sequencing methods. The first chapter introduces relevant background information, objectives of the study, and the structure of the thesis. After that, a comprehensive review of *de novo* peptide sequencing methods is given. It summarizes recent developments of computational methods for various experimental spectra, compares and analyzes their advantages and disadvantages, and points out some future research directions. Having these potential research directions, the thesis next presents two novel methods designed for higher-energy collisional dissociation (HCD) spectra and electron capture dissociation (ECD) (or electron transfer dissociation (ETD)) spectra, respectively. These methods apply new spectrum graph models with multiple types of edges, integrate amino acid combination (AAC) information and peptide tags, and consider spectrum-specific information to suit different spectra. After that, multiple spectra sequencing problem is studied. A framework for *de novo* peptide sequencing of multiple spectra is given with applications to two different spectra pairs. One pair is spectrally complementary to each other, and the other is similar spectra with property differences. These methods include effective spectra merging criteria and parent mass correction steps, and modify the previously proposed graph models to fit the merged spectra. Experiments on several experimental MS/MS spectra datasets and datasets pairs show the advantages of the proposed methods in terms of peptide sequencing accuracy. Finally, conclusions and future work directions are given at the end of the thesis.

To summarize the work in the thesis, a series of novel computational methods for *de novo* peptide sequencing are proposed. These methods target different types of MS/MS spectra and their combinations. Experiential results show the proposed methods are either better than competing methods that already exist, or fill gaps in the suite of currently available methods .

Acknowledgements

First, I would like to express my deepest appreciation to my supervisors Prof. Fang-Xiang Wu and Prof. Tony Kusalik. Throughout my study, they have given me constant encouragement and suggestions about my research topic and methods, and useful comments regarding to my academic writing. This thesis would not have been completed without their help.

I would also like to express my gratitude to other members of my advisory committee Dr. Mark Keil, Dr. Ian McQuillan, and Dr. Gopalan Selvaraj for their positive assistance and great advice throughout these years.

I would also like to thank my current and previous research group members Jinhong Shi, Bolin Chen, Lin Wu, Lizhi Liu, Zheng Yuan, Weiwei Fan, Jian Sun, Yichao Shen and Wenjun Lin for their help in both my life and research work. Their accompany throughout my whole time at the University of Saskatchewan makes my life joyful and lively.

I would like to specially thank all my family members for their continuous support and love. Their consisting encouragement guides me through all the difficulties in my PhD. study.

Finally, I would like to express my sincere acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC), University of Saskatchewan (UofS) and the China Scholarship Council (CSC) for the Financial supports during my pursuit of the PhD. degree.

To my fiance Randy Lin

Contents

List of Tables

List of Figures

List of Abbreviations

AAC	Amino acid composition
CAD	Collision-activated dissociation
CID	Collision-induced dissociation
Da	Dalton
DAG	Directed acyclic graph
ECD	Electron capture dissociation
ETD	Electron transfer dissociation
ExD	Electron capture dissociation and electron transfer dissociation
GMET	Graph with multiple edge types
HCD	Higher-energy collisional dissociation
HMM	Hidden Markov model
IM	Immonium ion
MS/MS	Tandem mass spectrometry
MS/MS spectrum	Tandem mass spectrum
m/z	Mass-to-charge ratio
OFF	offset frequency function
PTMs	Post-translational modifications

Chapter 1

Introduction

1.1 Background

Proteins are crucial entities of biological organisms. One important step to understand proteins is to know their sequences. The following are three main reasons of studying protein sequences. First, since a protein's sequence is unique, to at least some degree it helps distinguish proteins. Second, the sequence is the primary structure of a protein, so it is the basis of understanding the higher level structure and function of the protein. Finally, since most basic cellular processes are carried out by multiple proteins through molecular interactions, it helps the study of protein-protein interactions and molecular biology of proteins [3].

In the study of protein sequences, a typical strategy is to break them into smaller parts, which are peptides, and infer peptide sequences first instead. Therefore, peptide sequencing has become a prime concern in current proteomics [4].

Peptides are organic compounds consisting of two or more amino acids. Amino acids are molecules containing an amino group, a carboxyl group and a side chain that varies between different amino acids. The side chain is represented by R in Figure 1.1. It is the residue that distinguishes one amino acid from another. One nitrogen atom and two hydrogen atoms comprise the amino group ($-NH_2$), and one carbon atom, one oxygen atom and one hydroxyl (OH) constitute the carboxyl group ($-COOH$). The general structure of an amino acid is shown in Figure 1.1.

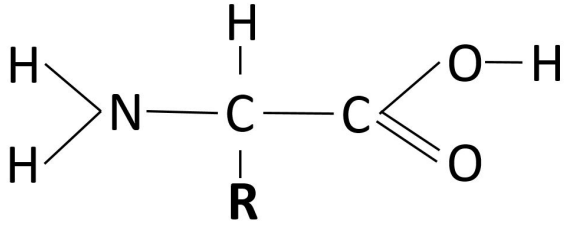


Figure 1.1: The general structure of an amino acid.

Two amino acids connect to each other when one's carboxyl group reacts with the other's amino group, creating a peptide bond ($-C(=O)NH-$) and losing a molecule of water (H_2O). The reaction of two amino acids is illustrated in Figure 1.2, where R and R' represent two side chains (residues). The elements in the

blue circle represent the water (H_2O) to be released, and the elements in the red circle represent the resulting peptide bond ($-C(=O)NH-$).

There are twenty standard amino acids in nature composing peptides and proteins. Each of them has been assigned an one-letter code for simplicity in use.

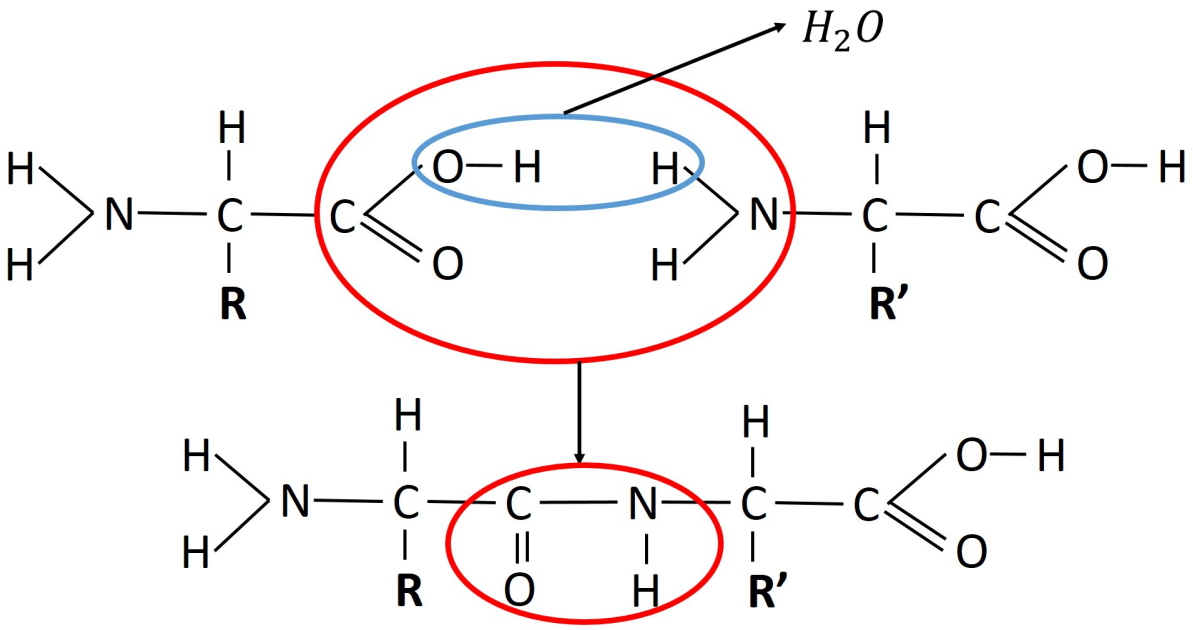


Figure 1.2: The reaction of two amino acids forming a peptide

Nowadays, tandem mass spectrometry (MS/MS) has emerged as a major technology for peptide sequencing because of the high throughput data it can generate in a short amount of time and its exceptional sensitivity [5–7]. A typical procedure of the MS/MS can be summarized as follows. Protein mixtures are first digested into suitably sized peptides for mass spectrometric analysis using site-specific proteases (usually trypsin). Then the peptides are ionized via an ionization process. After that, some selected ions are further fragmented into fragment ions, and the output of MS/MS is a diagram called a tandem mass spectrum (MS/MS spectrum) [8–11]. An MS/MS spectrum usually contains two kinds of information organized together as ordered pairs, the mass-to-charge ratio (m/z) values of fragment ions and the corresponding intensities. Therefore, MS/MS works like a charged sieve; we can only get a series of charged fragments from it [12]. Large molecules are broken into small pieces, and the problem of peptide sequencing is to find out the whole sequence of the peptide from these fragments [13].

Different kinds of fragment ions occur during MS/MS depending on the cleavage positions on peptide backbones, and the most frequently observed six kinds of them are named a-ions, b-ions, c-ions, x-ions, y-ions, and z-ions. With various fragmentation techniques used in MS/MS, different kinds of output spectra have differing dominant fragment ions and other unique properties. The commonly used ones include collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), electron capture dissociation (ECD)

and electron transfer dissociation (ETD). CID was the most commonly used technique when people started to use MS/MS for peptide sequencing, while the other three are newly developed techniques that are presently widely used. In this thesis, different kinds of MS/MS spectra are investigated and a series of computational peptide sequencing methods for various MS/MS spectra are developed.

1.2 Motivation and Objectives

When solving mass spectrometry-based peptide sequencing problems, database searching, peptide tagging and *de novo* sequencing are the most popular methods [14–18]. In database searching, theoretical spectra [19] are computed from an existing protein database and peptides are identified by matching the theoretical spectra to experimental spectra [20]. Peptide tagging [21, 22] produces partial sequences, often called tags, from an MS/MS spectrum, and then this information is usually combined with database searching or *de novo* sequencing to reduce the scale of computation. When combined with database searching, tags can be used as a filter to select candidate peptides that contain those partial sequences; when combined with *de novo* sequencing, a tag can be used as a starting point that is extended to get full candidate sequences. The last method, *de novo* sequencing, usually automatically interprets spectra using the masses of amino acids [1,23–26]. Because of its independence of databases, it can identify new proteins, proteins resulting from mutations, proteins with unexpected modifications and so on. Therefore, it is worth studying and developing the *de novo* sequencing methods.

Current *de novo* sequencing methods mainly have three limitations. First, the models constructed usually do not perfectly reflect MS/MS spectra. The second limitation is that less information is extracted than could be from an MS/MS spectrum. Lastly, early developed peptide identification methods only use one MS/MS spectrum, usually the CID spectrum, to infer the peptide sequence. The use of alternative MS/MS spectra and multiple spectra of the same peptide have the potential to significantly increase the accuracy and practicality of *de novo* sequencing. Facing these challenges and potentials, this study has the following objectives.

Objective 1: Review literature of currently developed *de novo* peptide sequencing methods and analyze their advantages and disadvantages.

Objective 2: Design a new model containing more useful information from MS/MS spectra for *de novo* peptide sequencing.

Objective 3: Design effective algorithms for *de novo* peptide sequencing using alternative spectra other than the traditional CID spectra.

Objective 4: Develop a framework of *de novo* peptide sequencing using multiple spectra based on graph-theoretic models.

1.3 Organization of the Thesis

This is a manuscript-style thesis. It is presented as a series of manuscripts for which I am the first author and primary investigator. These manuscripts were either published or submitted to specific journals during my Ph.D. study. The order of the manuscripts is accordant with the objectives. At the beginning of each chapter, detailed publication or submission information is given, followed by a brief introduction describing the connection of the manuscript to the context of the thesis. The format of all manuscripts has been adjusted to achieve consistency of format and style. All references of the publications have been unified and there is only one bibliography at the end of all chapters for the entire thesis.

The remainder of the thesis is organized as follows: Chapter 2 presents a comprehensive review of *de novo* peptide sequencing methods. It summarizes recent developments of computational methods for various types of typical experimental data, compares and analyzes their advantages and disadvantages, and points out some future research directions. Chapter 3 presents a new method named NovoHCD for *de novo* peptide sequencing. It applies a spectrum graph model with multiple types of edges, and integrates amino acid combination (AAC) information and peptide tags. This method has been applied to higher-energy collisional dissociation (HCD) spectra. It is compared to other similar methods on several experimental datasets to show its sequencing performance. Chapter 4 presents a *de novo* sequencing method for electron capture dissociation (ECD) and electron transfer dissociation (ETD) spectra named NovoExD. NovoExD modifies the model in NovoHCD to fit ECD/ETD spectra, and considers multiple peptide tags and fragment ion charge information. Its performance is then compared to another similar method. Chapter 5 presents a framework for multiple spectra sequencing and applies it to paired CID (or HCD) and ECD (or ETD) spectra. These spectra pairs have different dominant fragment ions and are complementary to each other. The performance of the framework is compared to another similar method named pNovo+. The results show that the proposed framework outperforms pNovo+ on several experimental datasets. Chapter 6 presents a *de novo* peptide sequencing method for CID and HCD spectra pairs. These spectra pairs have similar dominant ions but are accompanied by other ion types with different properties. Less attention has been paid in the literature to these spectra pairs. The proposed method includes a merging criteria of CID and HCD spectra and a parent mass correction step, and modifies a previously proposed algorithm for sequencing the merged spectra. The proposed method and other two methods designed for single spectrum (HCD and CID) sequencing are evaluated for performance. Finally, Chapter 7 concludes the thesis and gives possible future directions for this study. A full list of the publications produced during my Ph.D. studies is in Appendix A, and the copyright permissions of the published manuscripts included in this thesis are in Appendix B.

Chapter 2

Recent developments in computational methods for *de novo* peptide sequencing via tandem mass spectrometry (MS/MS)

Prepared as: Yan Yan, Anthony J. Kusalik and Fang-Xiang Wu. "Recent developments in computational methods for *de novo* peptide sequencing via tandem mass spectrometry (MS/MS)," Protein & Peptide Letters, accepted, August 2015.

The focus of the study in this thesis is MS/MS based *de novo* peptide sequencing methods. Before developing new computational methods for different kinds of MS/MS spectra, a comprehensive review of current *de novo* peptide sequencing methods is needed. CID was the most commonly used fragmentation technique when researchers started to apply *de novo* peptide sequencing methods to MS/MS spectra. Recently, with developments of fragmentation techniques, alternative MS/MS spectra are available. Among them, the most widely used ones are HCD, ECD, and ETD spectra. New *de novo* sequencing methods designed for these spectra have become available in recent years.

This chapter gives a review of recent developments of computational methods designed for various MS/MS spectra, especially methods for new, alternative spectra. With the availability of multiple types of spectra, methods designed for the use of spectra combinations have been developed, and they are reviewed in this chapter as well. This chapter summarizes different peptide sequencing methods available currently, compares and analyzes advantages and disadvantages of these methods, and points out potential future research directions. The review in this chapter provides a foundation for the study conducted in this thesis on development of new computational methods for *de novo* peptide sequencing.

Abstract

Tandem mass spectrometry (MS/MS) has emerged as a major technology for peptide sequencing. Typically, there are three kinds of methods for the peptide sequencing: database searching, peptide tagging, and *de novo* sequencing. *De novo* sequencing has drawn increasing attention because of its independence from existing protein databases and potential for identifying new proteins, proteins resulting from mutations, proteins with unexpected modifications and so on. Recently, with the improvements in the accuracy of MS/MS and development of alternative fragmentation modes of MS/MS, many new *de novo* sequencing methods have been formulated. This paper reviews these recently developed sequencing methods including those for alternative MS/MS spectra. The paper first introduces background knowledge on peptide sequencing and mass spectrometry, and then reviews *de novo* peptide sequencing methods for traditional CID spectra. After that, it focuses on the recent development of *de novo* peptide sequencing methods for alternative MS/MS spectra. In addition, methods using multiple spectra from the same peptide are surveyed. Finally, conclusions and some directions of future work are discussed.

2.1 Introduction

Peptide identification has become an important topic in current proteomics studies [27, 28]. Tandem mass spectrometry (MS/MS) has emerged as a major technology for peptide identification [3, 5, 6, 12]. Figure 2.1 shows a typical MS/MS experiment flowchart. In a typical MS/MS experiment, protein mixtures are first digested into suitably-sized peptides for analysis using site-specific proteases (usually trypsin). Then the peptides are ionized via an ionization process. The two most commonly used ionization sources are Matrix-Assisted Laser Desorption Ionization (MALDI) and Electrospray Ionization (ESI). More information about different ionization sources can be found in [2]. Peptide ions are then go through a mass analyzer and their mass-to-charge ratio (m/z) values and the corresponding intensities are measured. The generated data are called MS spectra. After that, some selected peptides (also called precursor ions) are further fragmented into fragment ions, and their tandem mass spectra (MS/MS spectra) are collected [8–11]. In this fragmentation process, different techniques may be applied and result in various kinds of MS/MS spectra. MS/MS spectra usually consist of two kinds of information, the mass-to-charge ratio (m/z) values of fragment ions and the corresponding intensities.

In MS/MS, parent peptide ions are fragmented into various kinds of fragment ions, mainly along the peptide backbone. The spine of a peptide contains three types of bonds ($C-C$, $C-N$, and $N-C$), and any of which may be broken in MS/MS. There are 6 types of ions that commonly occur, named a-ions, b-ions, c-ions, x-ions, y-ions, and z-ions according to their cleavage positions [29]. These ions can be in single or multiple charge states. Notations further indicate the positions at which the fragmentation of the peptide occurs. Figure 2.2 shows different cleavage sites and resultant ion types in detail. For example, in the cleavage

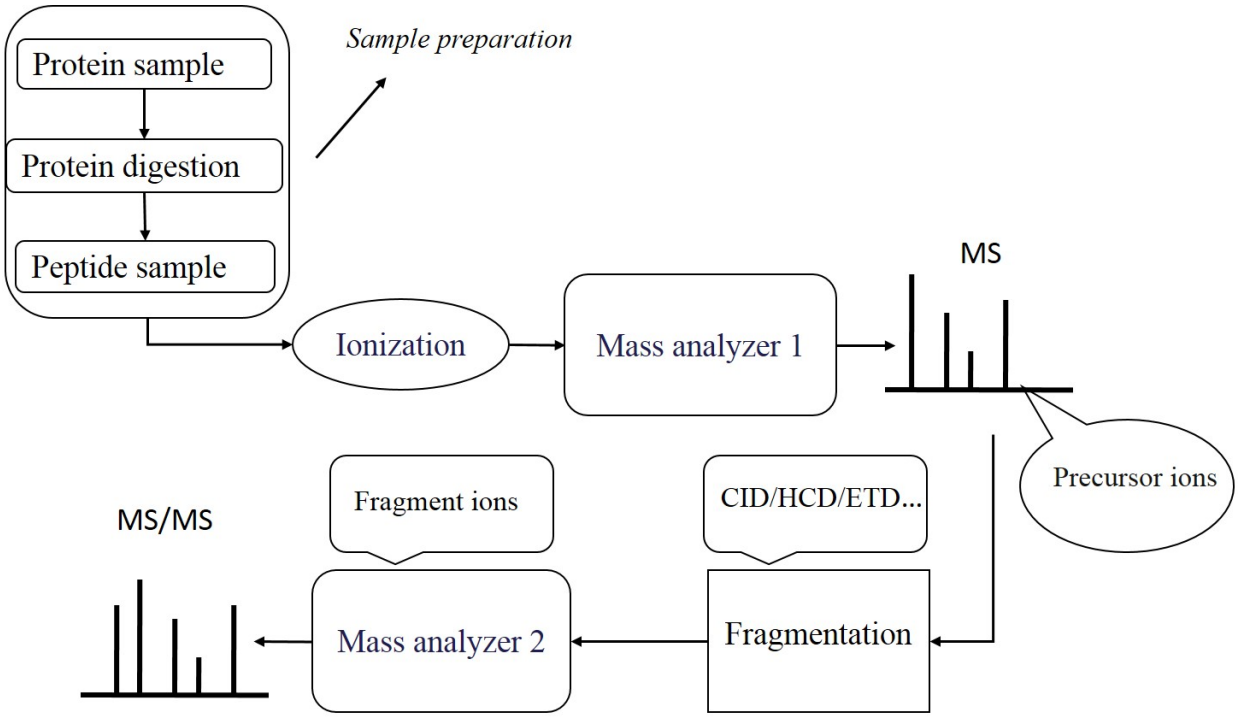


Figure 2.1: Flowchart of a typical MS/MS experiment

sites between residues R_1 and R_2, breaking the $C-C$ bond results in two complementary ions, an a_1-ion and an x_3-ion; breakage of the $C-N$ bond results in two complementary ions, a b_1-ion and a y_3-ion; and breaking the $N-C$ bond results in two complementary ions, a c_1-ion and a z_3-ion. In addition, the complementary ion relationships in Equations (1)-(3) hold for the six types of ions introduced above.

$$a_i + x_{N-i} = m_p + 2m_H, \tag{2.1}$$

$$b_i + y_{N-i} = m_p + 2m_H, \tag{2.2}$$

$$c_i + z_{N-i} = m_p + 2m_H, \tag{2.3}$$

where m_p is the mass of parent peptide P, N is peptide length, and $i \in \{1, 2, \ldots, N\}$.

Different fragmentation techniques in MS/MS yield differing dominant types of fragment ions. Collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD) yield b-ions and y-ions as dominating ions [30]. Electron capture dissociation (ECD) and electron transfer dissociation (ETD) preferentially produce variants of c-ions and z-ions, and occasionally a-ions [31–33]. In addition, all the fragment ions usually lose some small molecules such as H_2O and NH_3 during fragmentation [26, 34].

CID was the most commonly used fragmentation technique during the early development of peptide sequencing methods [14]. The principle of CID is shown in Figure 2.4. Selected precursor ions with positive charge values go into the collision cell that contains an inert collision gas, and fragment into product ions

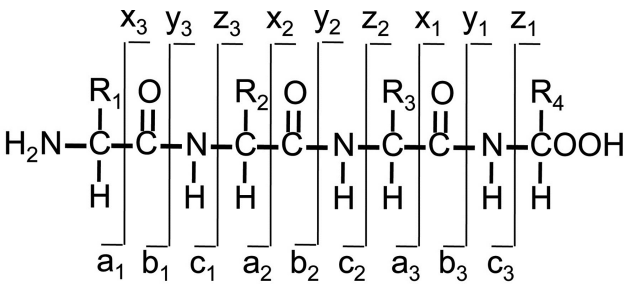


Figure 2.2: Different types of ions resulting from tandem mass spectrometry [1]

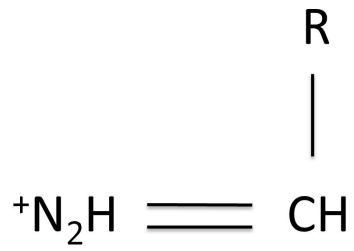


Figure 2.3: Structure of an immonium ion with residue R

and/or neutral losses. More information about the principles of different fragmentation techniques can be found in [2]. With the development of new techniques and instruments in recent years, alternative MS/MS spectra with new features have appeared [35]. HCD has similar dominating ions to CID but with more abundant ions in the low mass region (typically $\leq$ 200Da). Specifically, there are special types of ions shown on HCD spectra, the most informative ones being immonium ions (IMs) [36]. Figure 2.3 shows the structure of the IM. Other useful ions include b_1-ions, y_1-ions, and a_2/b_2-ion pairs.

ECD and ETD both produce variants of c-ions and z-ions, and the fragmentation process of these ions is shown in Figure 2.5. An ion with a dot "· means a radical fragment ion, which is a free radical species carrying a charge. ETD [37] is a modification of the ECD technique [38] which was designed for dissociation of multiply protonated peptide ions in MS/MS. In this study, we use ExD to represent ECD and ETD spectra as a whole. ExD produces high quality MS/MS spectra for multi-charged peptides and has no strong cleavage preferences. It utilizes a lower energy pathway than CID and HCD, thus preserving labile post-translational modifications (PTMs) [39–41]. All these features yield spectra containing useful information, and thus they have the potential to give satisfying peptide sequencing performance.

The goal of peptide identification is to infer the peptide sequence from an experimental spectrum. There are three main kinds of methods currently used for peptide sequencing: database searching, peptide tagging and *de novo* sequencing [42–45].

In database searching, theoretical spectra are computed from an existing protein database and peptides are identified by matching the theoretical spectra to experimental spectra. Different scoring schemes are employed to evaluate the matches [46]. The success of database searching thus relies on proper scoring functions [47] and the quality and extent of the existing databases. A major disadvantage of database searching is that it cannot

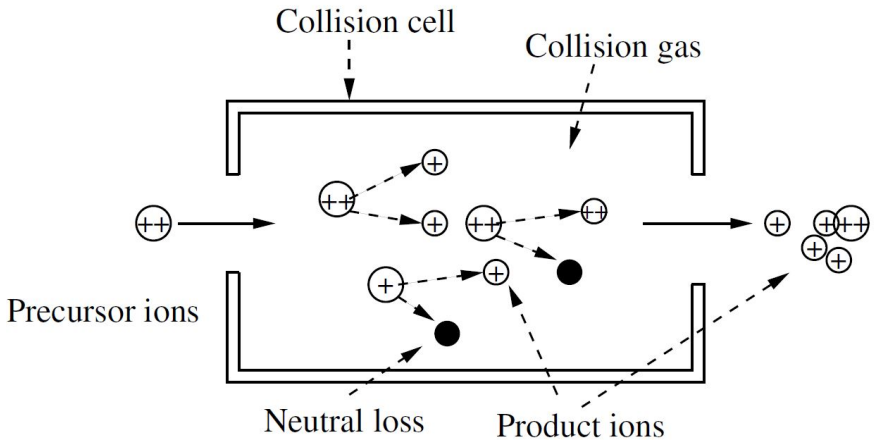


Figure 2.4: The principle of CID [2]

identify new or unknown peptides that are not included in the reference database. Peptide tagging [21, 22] usually produces partial sequences, often called tags, from an MS/MS spectrum, and then uses these tags to search against a protein database or to help with *de novo* sequencing. Since the method still searches in a known database, it cannot identify new or unknown peptides. The advantage of using tags is that it can dramatically reduce the search space and time needed in *de novo* peptide sequencing. *De novo* sequencing automatically interprets spectra using the masses of amino acids. It can identify new proteins, proteins resulting from mutations, proteins with unexpected modifications and so on. With the recent development of high mass-accuracy MS/MS and alternative fragmentation techniques, *de novo* sequencing has shown promising developments [48].

In this paper, we review recent developments in *de novo* peptide sequencing methods including some new methods using various types of spectra. The remainder of this review is organized as follows. Section 2 focuses on the peptide sequencing methods using the traditional CID spectra. Sections 3 summarizes some successful *de novo* peptide sequencing methods using alternative data such as HCD and ExD spectra. Section 4 introduces the multiple spectra sequencing problem and some recent solutions to it. Finally section 5 concludes the review and gives possible directions of future work.

2.2 Development of *de novo* peptide sequencing with the use of CID spectra

When researchers first started to apply *de novo* peptide sequencing methods to MS/MS data, CID spectra were commonly used. *De novo* sequencing started with straightforward methods like exhaustive listing and manual interpretation using unique features from CID spectra, and moved to refined and comprehensive methods like those using graph-theoretic models [49].

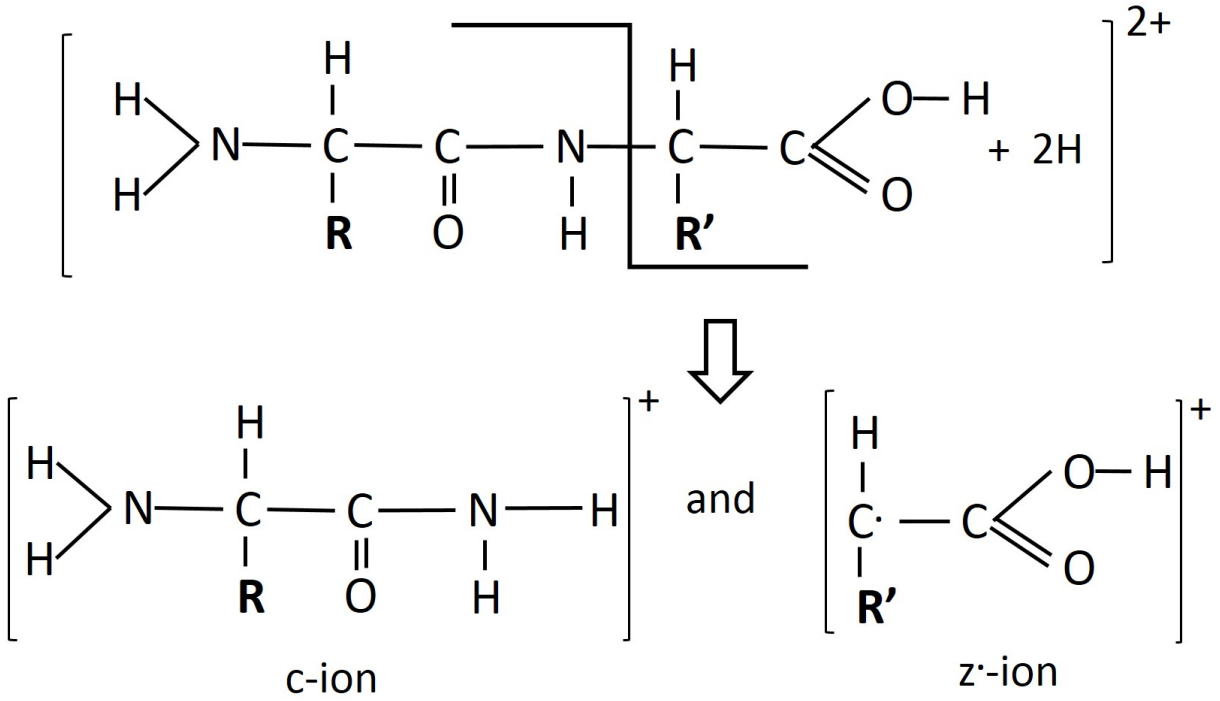


Figure 2.5: Fragmentation process leading to *c*- and *z*- type of ions in ExD

2.2.1 Early developments

Exhaustive methods involve listing all possible candidate peptide sequences that can result from a parent peptide mass in an MS/MS spectrum [50], and comparing them with an experimental spectrum to find the best match that has the highest possibility to generate such experimental spectrum. Typically, a suitably designed scoring scheme is involved to measure the similarity between the experimental spectrum and the theoretical spectrum of a candidate peptide. One computational difficulty with this approach is the exponential growth in the number of candidate sequences with increasing parent mass. Another challenge is the effect of inaccurate parent mass determination by the first stage of MS/MS, which leads to incorrect candidate sequences. Using these candidate sequences, results are poor even if the scoring scheme is fine. Manual interpretation usually employs graphical display of the MS/MS spectra, and relies on a human expert's visual skill to generate reliable results [51]. In this method, fragment ions that differ by the mass of one amino acid residue are highlighted by connected lines, thus allowing the visualization of ion series of the same type. This method may work well on simple spectra or spectra generated by short peptides. However, it requires a lot of human expertise and time.

The manual interpretation approach may seem quite simple and straightforward. Fortunately, there are ideas that can be used to enhance this approach to *de novo* peptide sequencing. For example, exponential growth in the number of candidate sequences limits the exhaustive listing for large peptides. For a peptide

with large mass value, it is quite time-consuming to get the comparison results for every possible candidate sequence with the experimental spectrum. But careful choice when listing partial peptide sequences can help with the sequencing of specific regions of a peptide. With the increased performance of computer hardware, this composition-based strategy has been applied in several recently proposed methods [52]. In addition, manual interpretation helps with finding new features in the experimental spectra, and provides ideas and prototypes for many methods based on the more recently developed graph-theoretic models. Fortunately, it is still quite worthwhile looking back and finding new ideas inspired by these initial approaches when designing new *de novo* sequencing methods.

2.2.2 Graph-theoretic methods

In *de novo* sequencing, graph-theoretic methods have proven to be quite successful and hence such algorithms have been widely used. In these methods, a tandem mass spectrum is typically represented as a graph called a spectrum graph. Each fragment ion is represented as a vertex and two vertices having the mass difference equal to one amino acid mass are connected by an edge [26, 49, 52].

The graph theoretical approach to *de novo* peptide sequencing was first proposed in the 1990's [53–55]. The main idea is to use a graph to represent different partial peptide masses and their relationships as interpreted from an MS/MS spectrum, and then to find suitable paths in the graph which indicate the likely peptide sequences that gave rise to the spectrum.

The Sherenga algorithm was proposed in 1999 [56] based on such a graph model. This method first identifies ion types in MS/MS data and then constructs the graph. This step helps with limiting the size of the graph. A merging approach is then applied using a greedy algorithm that determines which of the vertices in the spectrum graph can be merged into one vertex. The peptide sequencing problem is then transformed into a longest path-finding problem in a directed acyclic graph (DAG). Another useful strategy applied in the algorithm is a parent mass correction scheme since the accuracy of the parent mass is extremely important in *de novo* sequencing. A problem with this algorithm is that the best path might not correspond to a realistic solution because it may use multiple vertices associated with the same experimental spectral peak, which may be due to incorrect merging.

In order to solve some of the problems in the Sherenga algorithm, Chen *et al.* [57] proposed a dynamic programming approach to find the longest antisymmetric path in a spectrum graph. They introduced an NC-spectrum graph (NC denotes N-terminal and C-terminal, respectively) for a given tandem mass spectrum, which is constructed by assuming each peak in the spectrum is either a b-ion and a y-ion. Since each peak generates two vertices in the spectrum graph, the total number of vertices $2k + 2$, where k is the number of peaks in the spectrum. A dynamic programming algorithm is then applied to find longest paths from the graph. This idea was later developed into the *de novo* sequencing software named MSNovo [58].

In addition, some researchers applied other mathematical matrices to extract useful information from MS/MS spectra. One successful algorithm named PepNovo [59] uses a probabilistic network reflecting the

chemical and physical rules of the peptide fragmentation, and builds a spectrum graph to estimate the peptide sequence. This method performed well in *de novo* peptide sequencing, but is highly dependent on the training data used to generate the major parameters in the model, such as the fragmentation types and their probabilities. Thus, the accuracy of results varied with different MS/MS data [60].

2.2.3 Other methods

When solving the *de novo* peptide sequencing problem, some researchers develop their methods based on graph theoretical models, while others solve the problem in wholly different ways. One successful approach in the latter category uses a hidden Markov model (HMM), which is a statistical model describing sequential data with hidden information. The model ranks the likelihood between observed spectra and a given peptide sequence. The resultant software, NovoHMM [61], is based on this model. PEAKS [62], which is also not based on a spectrum graph model, uses dynamic programming to compute the best possible sequence among all possible amino acid combinations. It first applies a sophisticated dynamic programming algorithm [63] to output the 10,000 best candidate sequences, and then uses a carefully designed scoring scheme to rank them. Finally, it computes a confidence score for each sequence output. According to recent literature, PEAKS shows high sensitivity and good performance in peptide identification from MS/MS spectra [60]. In addition, there are many derivatives of the original PEAKS software described [31,64] that involve other new methods and techniques, thus making the integrated PEAKS package one of the most widely used software tools for *de novo* peptide sequencing.

2.2.4 Discussion

For *de novo* peptide sequencing using CID spectra, different algorithms and software have their comparative advantages and disadvantages, but in general, the trend is toward more effective peptide sequencing over time. However, one needs to keep in mind that the performance of each software package may vary among different spectra, instruments, techniques and experiment samples. Therefore, it is impossible to say which one is best for peptide sequencing or which surpasses all others. Instead, researchers should carefully choose software to fit the specific peptide sequencing problem they are facing.

There are still some limitations and problems in the above methods. The first one is that the spectrum is typically noisy. To solve this problem, preprocessing and denoising approaches are needed [65], such as the preprocessing in PEAKS and the work done by Sridhara *et al.* [66]. In addition, because the ions used to find paths should be of the same ion type with de novo sequencing, ion type separations are helpful in the subsequent peptide sequencing results. Some researchers have used a graph-theoretic approach to separate *b*- and *y*- ions in an MS/MS spectrum [67], but little further work has been done in this vein.

The second problem is missing data from MS/MS spectra. Researchers have tried to extrapolate as much useful information as possible from the models constructed, but they are typically not able to satisfactorily reflect the actual MS/MS spectrum, and there is still a need to generate more accurate information from the

outset [68].

There is still the potential to develop better strategies and newer models to improve sequencing performance. For example, MSNovo [58] has integrated a new probabilistic scoring function with a mass array-based dynamic programming algorithm to predict peptide sequences. In addition, researchers can change the definitions of vertices and edges in current spectrum graph models and investigates more relationships among ions (not just the mass difference of amino acids [69]). Since the spectrum itself contains limited information, another possible way is to combine other approaches — for example, database searching — to estimate the best peptide sequence from the experimental spectrum. Tag-based approaches following this strategy, such as PARPST [70] and Vonode [69], are widely used. Instead of inferring the whole sequence, these algorithms first output the best peptide tag with a strict error tolerance and then use the tag to search potential peptides in a previously specified database. These algorithms usually have advantages of shorter computation time compared to traditional database searching methods and more accurate peptide sequencing results compared to the traditional *de novo* peptide sequencing. Therefore, tag-based methods are worth studying deeper in a deeper fashion in the future.

2.3 *De novo* peptide sequencing with alternative MS/MS spectra

CID was the most commonly used fragmentation technique when researchers started to use tandem mass spectrometry for peptide sequencing. In recent years, alternative MS/MS spectra such as HCD and ExD have appeared with the development of new techniques and instruments. Our study of these new spectra starts with the investigation of their properties and unique features that could be used for peptide sequencing, and then moves to the development of computational methods. In particular, we review the development of *de novo* peptide sequencing using HCD and ExD spectra.

2.3.1 Methods using HCD spectra

HCD produces similar dominant ions to the traditional CID technique, but generates more ions in the low mass region (typically $\leq$ 200Da). This includes ions specific to HCD spectra such as IMs (immonium ions), which are introduced and studied in [71]. Another systematic study of the properties of HCD spectra can be seen in [36]. From these studies, researchers have learned how to use this kind of spectra to infer peptide sequences using *de novo* methods.

One successful method specifically developed for HCD spectra is pNovo [72]. It applies a spectrum graph model and combines IMs and internal fragment ion information from HCD spectra, and has achieved superior peptide sequencing results. Since there are few algorithms available designed specifically for HCD spectra and CID spectra have similar dominant ions, the authors compared the performance of pNovo with other algorithms designed for CID spectra. The results on various test data showed that pNovo has higher sequencing accuracy than two previous algorithms, PEAKS [62] and PepNovo [59].

Although *de novo* peptide sequencing for HCD spectra takes advantage of advanced instruments and specifically designed methods, there are still limitations to these methods, especially the insufficient information extracted from the spectrum graph model. For example, in a traditional spectrum graph like the one used in pNovo [72], an edge is drawn between two vertices only when the mass difference between the two vertices is equal (or close) to one of the 20 amino acid masses. Thus if there are peaks missing in the spectrum or the m/z values are inaccurate, it could be very difficult to find a path through the graph representing the correct peptide sequence. Therefore, a more suitable model with more information included is very appealing. Since the traditional model only considers amino acid mass difference, a graph model from MS/MS spectra that includes multiple relationships is expected to have better performance [21].

NovoHCD [73] is such a method that uses a modified spectrum graph model with multiple types of edges (called a multi-edge graph) for peptide sequencing. It also combines amino acid combinations (AACs) and peptide tags to make the problem easier to solve. This method first uses peptide tags to separate a whole peptide sequence into three parts: prefix, tag, and suffix. It then builds multi-edge graph models on the prefix and suffix separately for sequence interpretation, and uses AAC information to limit the number of edges in the graph. It finally combines these three parts to generate complete sequences of candidate peptides. Immonium ions observed particularly in HCD spectra are used in the candidate peptide ranking of NovoHCD. From experiments on five HCD spectra datasets, NovoHCD outperforms pNovo in terms of the accuracy [73].

2.3.2 Methods using ExD spectra

When it comes to the other new types of spectra, ExD is a popular one because of its unique properties. ExD is a new technology that has properties different from CID and HCD. Recent studies of MS/MS spectra produced by ExD have focused on the characteristics of the spectra [32, 33, 39], how various PTMs can be identified from the spectra [40], and performance of peptide sequencing methods using such spectra [17, 74].

Some of the methods that use multiple spectra for sequencing include an option to use ExD spectra alone; for example, the one introduced in [75]. These methods typically consider only a subset of the features in ExD spectra because their focus is not ExD spectra exclusively. Up to now, little attention has been paid to *de novo* sequencing methods using solely ExD spectra. However, one recently developed method, NovoExD [17], takes advantages of the unique features of ExD spectra to infer peptide sequences from them alone. NovoExD uses a novel spectrum graph model, considers multiple peptide tags to separate a peptide into small mass regions, and integrates fragment ion charge and amino acid composition (AAC) information. It combines small regions to output complete sequences of candidate peptides. In addition, a charge determination step is used to extract more information from highly charged ExD spectra. Experiments on three ExD datasets show an improvement of over 20% in average full length peptide sequencing accuracy for NovoExD as compared to a multiple spectra method with an option for ExD spectra. Given the limited number of methods for ExD alone and the unique features of such spectra, improved methods for them would be useful and worth studying.

2.3.3 Discussion

The two methods NovoHCD and NovoExD achieved superior experimental results compared to other similar approaches. They have a similar framework but with modifications suitable for the new types of spectra they utilize (HCD and ExD, respectively). The reasons for their success can be summarized as the following. First, the framework replaces the traditional spectrum graph with a multi-edge graph, thus including more relationships between pairs of peaks in a spectrum. Secondly, it separates whole peptide sequences into smaller parts, and solves each sequencing subproblem separately. Lastly, AAC information is effectively incorporated into the peptide sequencing process to limit the number of edges in the generated graph models, which also reduces the computational cost for these methods.

In the future, exploring the unique properties of these new types of spectra, and designing innovative techniques and models that utilize the novel types of information they contain, are potential research directions that could further enhance *de novo* sequencing performance.

2.4 Multiple spectra *de novo* peptide problem and its development

De novo peptide sequencing methods have existed for several decades, but there are still challenges in correctly identifying peptide sequences. Traditional *de novo* peptide sequencing methods use only one MS/MS spectrum to conduct peptide sequencing. The main shortcoming of these approaches is the limited fragmentation information extracted from only one experimental spectrum. Therefore, identification accuracy is limited. Apart from using new and high quality spectra for sequencing, there is another way to improve performance, which is to combine multiple spectra from the same peptide but from different technologies to conduct *de novo* sequencing [74]. For instance, CID (and/or HCD) and ExD spectra belonging to the same precursor can be paired to obtain more fragmentation information for peptide sequencing [75, 76]. The use of a pair of spectra is the major focus of multiple spectra peptide sequencing.

In CID fragmentation, b-ions and y-ions are the primary results from cleavage of the peptide bond, while ExD fragmentation has no preferred cleavage sites except proline, which leads to a more uniform distribution of the fragment ions, mainly consisting of c-ions and z-ions. In addition, CID fragmentation tends to produce more fragment ions from cleavages in the middle of the sequence while ExD fragmentation prefers cleavages at the end of the sequence [77]. This makes CID and ETD perfectly complementary techniques and leads to a more comprehensive coverage of a peptide sequence with fragment ions [78].

In the following, some major developments of utilizing multiple spectra for *de novo* peptide sequencing are discussed and compared. Although the current trend is to use a pair of spectra, some of the methods introduced below have the potential or are already able to utilize three or more spectra.

To our knowledge, the earliest method designed for multiple spectra sequencing was the one developed

by Savitski *et al.* [79]. This method utilizes spectra of the same peptide from collision activated dissociation (CAD) and electron-capture dissociation (ECD) together. Their algorithm first uses ions that have supporting ions in the other spectrum to create a backbone of the sequence, then uses complementary ion pairs and other ions from the two spectra to extend the sequence until a full sequence is obtained. This is a linear greedy algorithm that can be readily computed.

Other methods that fall into this category include CompNovo [77], Spectrum Fusion [80], pNovo+ [75], and NovoPair [81]. CompNovo [77] employs a divide-and-conquer approach combined with a mass decomposition algorithm. It uses informative ions from CID spectra to separate the whole spectra into suitably-sized segments controlled by the divide-and-conquer criteria. It lists all possible amino acid compositions (AACs) and compares them with the two experimental spectra to infer the peptide sequence that generated the spectra. Spectrum Fusion [80] is proposed as a generic algorithm for multiple spectra sequencing. It combines multiple spectra into a synthetic spectrum, and conducts *de novo* sequencing using the synthetic spectrum. The method was evaluated on CID and ETD spectra pairs [80], but the method framework is applicable for more and other types of spectra. In addition, this method includes a parent peptide correction strategy, which is essential in solving multiple spectra sequencing problem.

CompNovo and Spectrum Fusion are more suitable for CID and ExD spectra because of the features they consider, while the latter two, pNovo+ and NovoPair, are designed for HCD and ExD spectra. pNovo+ first uses an offset frequency function (OFF) from Sherenga [56] to learn all ion types considered in the algorithm, and then uses isotopic clusters to determine ion charge and convert all ions to +1 form to simplify subsequent calculations. After this, the algorithm builds a spectrum graph and finds k longest paths (keeping just the k longest paths from 0 to the current vertex every time), inferring the peptide sequences from them. NovoPair uses a new spectrum graph model (GMET) developed from [73] and considers different fragment ion types occurring in both spectra. It uses length-three peptide tags to separate a peptide into small regions, and integrates amino acid composition (AAC) information into the graph model.

Apart from combining information from multiple spectra from the same subsequent peptide for peptide sequencing, there are other ways of utilizing the information. He and Ma [31] presented the algorithm ADEPTS that includes a new scoring function to select the best peptide candidates from the sequencing results output by traditional methods; for example, candidates produced by the PEAKS [62] software for each of two spectra. This method does not try to combine information from multiple spectra to design a new sequencing method, but rather focuses on a new scoring function designed to select the correct sequences from all candidates. The hypothesis of ADEPTS is that the correct sequence is already generated by single-spectrum sequencing methods, and the only need is a good selection scheme to find that correct one. Therefore, ADEPTS is expected to have good performance for the case that at least one of the multiple spectra is of good quality. ADEPTS has implemented a new way of using multiple spectra for *de novo* peptide sequencing that integrates conveniently with current peptide sequencing algorithms. It also has the potential to be extended to the case of more than two spectra.

There are other studies that use multiple spectra, and we just give a brief introduction here. Altelaar *et al.* [82] utilized a special enzyme to break the peptide backbone at the N-terminal side of lysine residues, and combined CID and ETD spectra to conduct peptide sequencing. Guthals *et al.* [83] used MS/MS triplets (CID/HCD/ETD) from overlapping peptides produced by different enzymes to infer peptide or even protein sequences. The longest sequence from their experiments is up to 200 amino acids. Shen *et al.* [74] compared the peptide sequencing performance of different methods when using various types of spectra solely or together. Recently, Jeong *et al.* [84] proposed a universal *de novo* sequencing tool based on spectrum graph model and claim it can be used for various types of spectra.

2.4.1 Discussion

The use of multiple spectra has become popular in the past decade and major attention has been paid to the use of a pair of spectra from the same peptide. There are still a limited number of methods available for this problem, and often it is hard to determine the superiority among them. Both CompNovo [77] and Spectrum Fusion [80] have been shown to achieve better peptide sequencing results on various experimental datasets compared to those methods that only use a single spectrum such as PEAKS [62] and PepNovo [59]. ADEPTS has shown better performance compared to both PEAKS and CompNovo [31]. NovoPair claims to outperform pNovo+ in terms of full length peptide sequencing accuracy on three different experimental dataset pairs [81]. It is important to notice that the above conclusions are based on different experimental data used in each study, and due to differences in properties of the data, these comparisons are not necessarily comprehensive and conclusive.

Despite the limitations, all the introduced multiple spectra sequencing methods have opened a new door for *de novo* peptide sequencing and provide a promising way to solve some of the current challenges facing traditional *de novo* peptide sequencing methods. Since most of the recent methods are mainly focused on a pair of spectra, the extension of these algorithms to more spectra can be a major focus for future study. When more types of data are available with improvements in fragmentation technique, we should be able to apply the current algorithms to them with proper modifications. For example, the algorithm developed by He and Ma [31] can be easily applied to more spectra with a small change of scoring function. In addition, more spectrum-specific features should be investigated and applied in the methods using multiple spectra. Currently, some methods mainly use mass-based features [77, 79] and some use intensity-based features [31]. These features are usually used to extract information from all spectrum types. They are general features common to all spectra and not spectrum-specific features. By leveraging spectrum-specific features, such as IMs and internal fragments in HCD spectra, we may be able to extract more information from a spectrum and achieve better sequencing results. Finally, since the use of pairs of MS/MS spectra have been applied to database search [85,86], combination approaches of *de novo* sequencing and database searching using multiple spectra are promising as a future study direction.

2.5 Conclusions and outlook

De novo peptide sequencing has evolved over several decades and numerous methods have been published. This paper reviewed recent developments in the area, especially methods developed for new types of MS/MS spectra. The paper first introduced background knowledge on peptide sequencing and principles of the experimental instrument, and then reviewed *de novo* peptide sequencing methods for traditional CID spectra. After that, it summarized recent developments in *de novo* sequencing using alternative spectra, with the focus on the methods using a single spectrum and multiple spectra from the same peptide.

Different algorithms and software have their respective advantages and disadvantages, and it is unfair to say that one categorically surpasses all others. At the same time, there are still problems and limitations in current methods [87]. In light of the problems, the following are some future work directions that we believe would be worth considering. Firstly, since an MS/MS spectrum is noisy and with missing data, improved preprocessing methods applied before peptide sequencing would be quite useful [88–90]; for example, preprocessing methods for alternative MS/MS spectra. Secondly, since some algorithms are highly dependent on an accurate parent peptide mass, more strategies can be used to improve that mass accuracy. Thirdly, in the use of m/z ratios, many current algorithms have not considered various charge states (differing values of z). Since multiple charges occur often in MS/MS, especially in ExD spectra [17], charge determination methods would be quite useful to assist with *de novo* peptide sequencing [91–93]. Fourthly, for the multiple spectra case, more efforts can be put toward new method development and performance comparisons among different methods. Finally, for the combination of different methods, more effective ways can be devised to combine *de novo* peptide sequencing with database searching; for example, Zhang *et al.* [64] give a new way to combine the two kinds of methods. Also, some tag-based methods have been applied in the combination of *de novo* peptide sequencing and database searching [69,70,73]. Researchers can work on multiple tag generation in these methods and cleavage determination of those tags, both of which could improve the performance of this combination of methods.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Competing interests

The authors declare that they have no competing interests.

Chapter 3

NovoHCD: *De novo* peptide sequencing from HCD spectra

Published as: Yan Yan, Anthony J. Kusalik and Fang-Xiang Wu. "NovoHCD: *De novo* peptide sequencing from HCD spectra," IEEE Transactions on NanoBioscience, vol.13, no.2, pp.65-72, June 2014.

In the previous chapter, a comprehensive review of *de novo* peptide sequencing methods for MS/MS spectra is given. Findings and analysis in it point out limitations and gaps in current methods available and suggest potential directions for improvements and new method developments. Therefore, in this chapter, a study of a new *de novo* peptide sequencing method for HCD spectra is presented.

It is mentioned in the previous chapter that with the development of MS/MS fragmentation techniques, new types of MS/MS spectra are available and computational methods designed for them have been developed. However, current methods often fail to build satisfying models that efficiently extract information from HCD spectra. New models that are suitable for HCD spectra better are needed.

In this chapter, a *de novo* sequencing method including a new type of spectrum graph model is presented. This graph includes multiple types of edges representing different kinds relationships between two ions in a spectrum. Amino acid composition information and peptide tags are used in this method to limit the size of the graph and simplify the calculation in the peptide sequencing. Experimental results on several HCD spectral datasets show that the proposed method, NovoHCD, outperforms other competing methods in terms of full length sequencing accuracy.

Abstract

In recent years, *de novo* peptide sequencing from mass spectrometry data has developed as one of the major peptide identification methods with the emergence of new instruments and advanced computational methods. However, there are still limitations to this method; for example, the typically used spectrum graph model cannot represent all the information and relationships inherent in tandem mass spectra (MS/MS spectra). Here, we present a new method named NovoHCD which applies a spectrum graph model with multiple types of edges (called a multi-edge graph), and integrates into it amino acid combination (AAC) information and peptide tags. In addition, information on immonium ions observed particularly in higher-energy collisional dissociation (HCD) spectra is incorporated. Comparisons between NovoHCD and another successful *de novo* peptide sequencing method for HCD spectra, pNovo, were performed. Experiments were conducted on five HCD spectral datasets. Results show that NovoHCD outperforms pNovo in terms of full length peptide identification accuracy; specifically, the accuracy increases 13%–21% over the five datasets.

3.1 Introduction

There is growing interest in the identification of peptide sequences on the proteome-wide scale. Tandem mass spectrometry (MS/MS) has emerged as a major technology for peptide identification [5,6]. In a typical MS/MS experiment, protein mixtures are first digested into suitably sized peptides for mass spectrometric analysis using site-specific proteases (usually trypsin). Then the peptides are ionized via an ionization process. After that, selected peptides are further broken into fragment ions, and their tandem mass spectra are collected [8–11].

In MS/MS, peptide ions are fragmented into various kinds of fragment ions, named a-, b-, c-, x-, y-, and z-ions. Different fragmentation techniques used in MS/MS yield differing dominating types of fragment ions. Collision-induced dissociation (CID) is one of the most commonly used fragmentation techniques, and it yields b-ions and y-ions as dominating ions. Higher-energy collisional dissociation (HCD) has similar dominating ions to CID but with more abundant ions in the low mass region (typically $\leq$ 200Da). Specifically, there are special types of ions shown on HCD spectra, and the most informative ones are immonium ions (IMs) [36]. Other useful ions include b_1-ions, y_1-ions, and a_2/b_2-ion pairs. There are also other types of fragmentation techniques, such as electron capture dissociation (ECD) and electron transfer dissociation (ETD). These preferentially produce variants of c-ions and z-ions, and occasionally a-ions, and thus can be viewed as the complement of CID or HCD [31].

There are two kinds of methods widely used for peptide sequence identification from MS/MS data: database searching and *de novo* sequencing. In database searching, theoretical spectra are computed from an existing protein database and peptides are identified by matching the theoretical spectra to experimental spectra. The major disadvantage of database searching is that it cannot identify new, currently unknown

peptides since a prior database is always needed. *De novo* sequencing, on the other hand, estimates peptide sequences without the help of a database; it automatically interprets spectra using the masses of amino acids. Therefore, this method can identify new proteins, proteins resulting from mutations, proteins with unexpected modifications and so on. During recent years, especially with the development of high mass accuracy MS/MS, *de novo* sequencing has drawn increasing attention [48]. Therefore, this study focuses on the improvement of *de novo* peptide sequencing.

In *de novo* peptide sequencing, spectrum graph modeling has proven to be quite successful and hence has been widely used [21, 52, 59, 62, 72]. In this approach, a tandem mass spectrum is typically represented as a graph. Each fragment ion, corresponding to a peak in a spectrum, is represented as a vertex and two vertices having a mass difference equal to one amino acid mass are connected by an edge. The main idea of this method is to find paths in the graph that represent peptide sequences potentially giving rise to the spectrum.

There is another method used in the peptide sequencing called sequence tagging that comprises a middle path between database searching and *de novo* sequencing [21, 22]. Peptide tagging first outputs partial sequences, usually called tags, from a MS/MS spectrum, and then uses these tags to search against a protein database to interpret the spectrum. The use of tags can dramatically reduce the search space and time needed, which makes it popular for some peptide identification problems. In addition, peptide tags also have the potential to help *de novo* peptide sequencing since it provides useful information of partial peptide sequences.

With the appearance of alternative MS/MS data resulting from different fragmentation techniques, novel computational methods have emerged to enhance *de novo* peptide sequencing performance [94]. pNovo [72], which applies a spectrum graph model and combines IM and internal fragment ion information from HCD spectra, has achieved superior peptide sequencing results. Its performance on various testing data has been shown to be better than that of two previous algorithms, PEAKS [62] and PepNovo [59]. Other research has focused on the ability of pre-processing and charge determination in ETD spectra to improve sequencing results [66]. Another popular approach is combining different types of spectra from the same peptide to achieve better results [74]. For instance, CID (and/or HCD) and ETD (or ECD) spectra belonging to the same precursor can be paired to obtain more fragmentation information for peptide sequencing [75, 76].

Although *de novo* peptide sequencing has improved with the advance of instruments and computation, it still has shortcomings, especially the limited amount of information extracted from the spectrum graph model. For example, in a traditional spectrum graph, an edge is drawn between two vertices only when the mass difference between the two vertices is equal (or close) to one of the 20 amino acid masses. Then if there are peaks missing in the spectrum or the m/z values are inaccurate, it could be very difficult to find a path from the graph representing the correct peptide sequence. Therefore, a more suitable model with more information included is very appealing. Since the traditional model only considers amino acid mass difference, a multi-edge graph that includes multiple relationships from MS/MS spectra is expected to have

better performance [21].

Natural choices for extracting more information from a spectrum are peptide tags and amino acid combinations (AACs). Peptide tags provide information about partial peptide sequences, information that would definitely be useful to *de novo* peptide sequencing. AAC, which consists of order-independent amino acid composition for a peptide, is also quite helpful [95, 96], especially in limiting the edges of a spectrum graph. In general, graph-based algorithms compute the mass difference between the two vertices and compare it with the theoretical residue mass of the 20 amino acids. However, in practice, a peptide usually does not contain all 20 amino acids. In addition, due to the high noise level of the MS/MS spectra, wrong matches of amino acids can easily happen if we compare with all 20 possible masses. When the correct AAC is known, instead of considering all 20 amino acids, only those in the AAC set are considered when determining whether there is an edge between two vertices. Therefore, identifying the proper AACs can ultimately reduce computational time and eliminate false positive edges in a spectrum graph.

In this paper, we present a new method, NovoHCD, which uses a modified spectrum graph model with multiple types of edges (called a multi-edge graph) for peptide sequencing. It is developed from [21, 94] and combines amino acid combinations (AACs) and peptide tags to infer peptide sequences. The remainder of this paper is organized as follows. Section 2 presents the proposed *de novo* sequencing method. Section 3 describes performance experiments and results. Finally Section 4 concludes the paper and gives future work.

3.2 Methods

The proposed method, NovoHCD, first uses peptide tags to break the whole peptide sequencing problem into three parts: sequencing of tags, prefixes, and suffixes of the peptide. That is, the peptide sequence being inferred is separated into three parts (prefix, tag, and suffix). The method builds separate multi-edge graph models for each prefix and suffix, and AAC information is used to limit edges of each graph. This method then combines these three parts to output complete sequences of candidate peptides. In addition, immonium ions observed particularly in HCD spectra are used in the candidate peptide ranking step. The method flow chart is shown in Figure 3.1.

3.2.1 Multi-edge graph model

In the new multi-edge graph $G = (V, E)$, each peak (corresponding to a fragment ion) in the experimental spectrum is represented as a vertex v and its mass to charge (m/z) value is denoted as m_v. In the following, unless otherwise specified, we assume that the fragment ions have $z = 1$. A is taken to be the set of 20 amino acids, and $a_i \in A$ is a certain amino acid. a_i is also used to represent its residue mass. m_{loss} is defined to be the mass of some small molecules lost from fragment ions, which typically include H_2O, NH_3, CO- and NH- groups. For a string of amino acids $P = a_1 a_2 \dots a_n$, define $|P|$ to be the mass sum of all amino acids in P. Thus $|P| = \sum_{i=1}^{n} a_i$. Then the mass of parent peptide P (denoted as m_P) becomes $m_P = |P| + m_{H_2O}$,

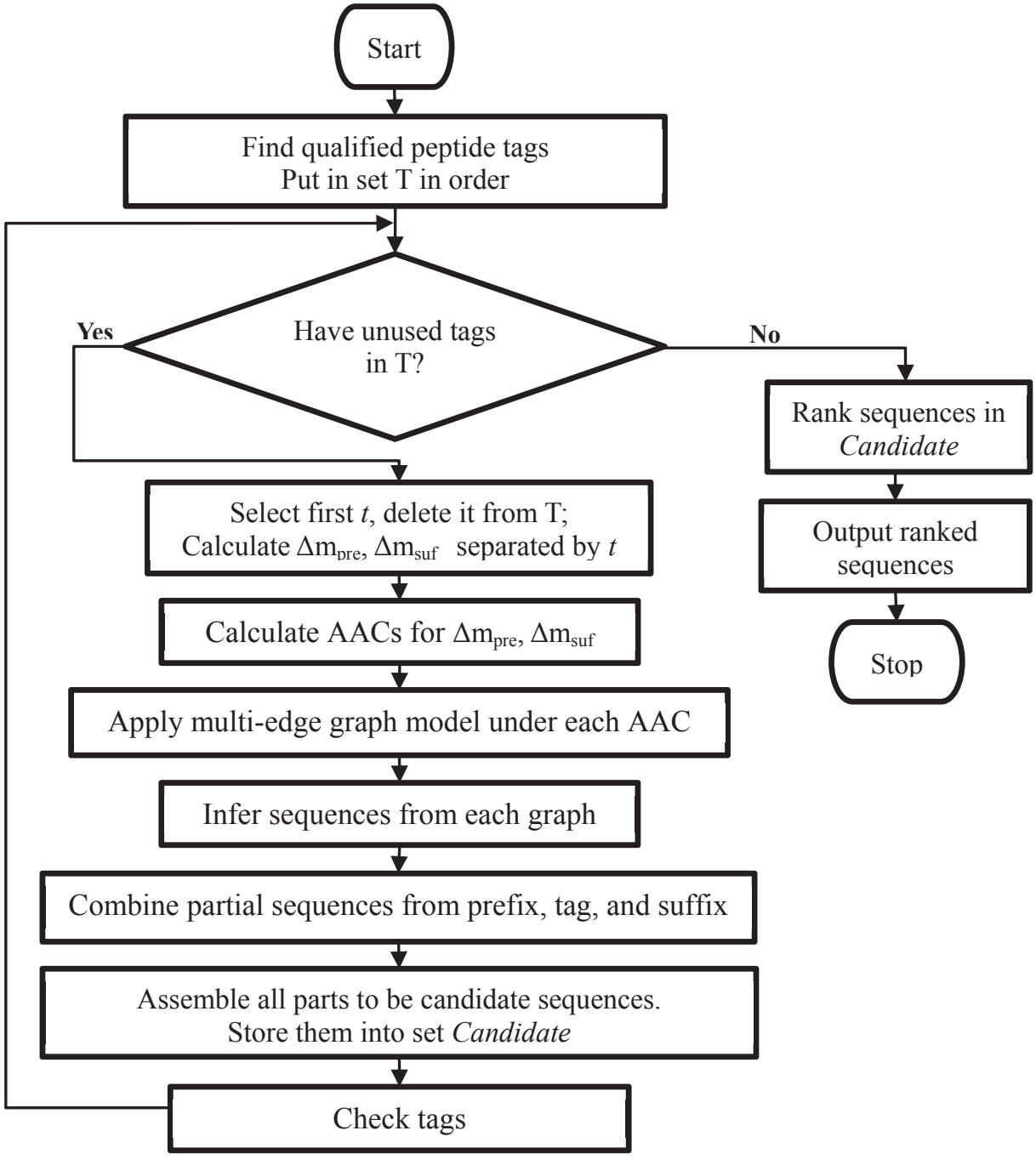


Figure 3.1: Method flow chart. All peptide tags are stored in set T, and t represents a tag in T; Δm_{pre} and Δm_{suf} represent the mass values of the prefix and suffix separated by t, respectively.

and the mass sum of a pair of complementary ions (charge +1) derived from P, denoted as M, becomes $M = m_P + 2$. The above notions are similar to those used in [63]. $\forall u, v \in V$, the following types of edges are considered in G.

1. Type 1 edge, e^1_{uv}: A directed edge e^1_{uv} is drawn from u to v when $\mid (m_v - m_u) - a_i \mid \leq \theta_1$, where θ_1 is a given threshold. We label e^1_{uv} with a_i. Here, vertex u and v are expected to have a mass difference of a single amino acid, a_i. This is the type of edge defined in a traditional spectrum graph.

2. Type 2 edge, e^2_{uv}: An undirected edge e^2_{uv} is drawn between u and v when $\mid M - (m_v + m_u) \mid \leq \theta_2$, where θ_2 is a given threshold. Here u and v are viewed as complementary ions.

3. Type 3 edge, e^3_{uv}: An undirected edge e^3_{uv} is drawn between u and v when $\mid (m_v + m_u) - M - a_i \mid \leq \theta_3$, where θ_3 is a given threshold. We label e^3_{uv} with a_i.

4. Type 4 edge, e^4_{uv}: An undirected edge e^4_{uv} is drawn between u and v when $\mid M - (m_v + m_u) - a_i \mid \leq \theta_4$, where θ_4 is a given threshold. We label e^4_{uv} with a_i.

5. Type 5 edge, e^5_{uv}: A directed edge e^5_{uv} is drawn from u to v when $\mid (m_v - m_u) - m_{loss} \mid \leq \theta_5$, where θ_5 is a given threshold. Here u and v are different ions from the same cleavage site of the peptide.

All θ values mentioned above are specified by the user. For the HCD spectra used in the experiments of this paper, all θ values were set to be 0.01Da. Figure 3.2 shows (part of) a multi-edge graph. In this graph, vertices u and v have a mass difference of a_i, and v and t have a mass difference of a_j. u_c, v_c and t_c are complementary ions of u, v and t, respectively. u_{loss}, v_{loss}, t_{loss}, u_{c-loss}, v_{c-loss} and t_{c-loss} represent any loss of small molecules from u, v, t, u_c, v_c and t_c, respectively. In this graph, Type 1 edges are black with arrows, Type 2 edges are blue, Type 3 edges and Type 4 edges are purple and green, respectively, with arrows and Type 5 edges are red with arrows. One can see that, except for Type 2 and Type 5 edges, all other edges have amino acids labeling them, and one amino acid can be inferred through all vertices from the same cleavage site of the peptide. For example, in Figure 3.2, amino acid a_i can be inferred from the vertex set $VS_{uv} = \{u, v, u_c, v_c, u_{loss}, v_{loss}, u_{c-loss}, v_{c-loss}\}$. The vertices $u_{loss}, v_{loss}, u_{c-loss}, v_{c-loss}$ represent just one kind of small neutral loss, and in real experiments, different kinds of losses may occur and there would be more vertices representing small neutral losses in the multi-edge graph.

Here, we define the graph induced by all vertices from the same cleavage site of the peptide as a *basic structure* of the multi-edge graph G, denoted $G[VS]$. From each $G[VS]$, an amino acid a_g is expected to be inferred. Therefore, we add this amino acid into the notion of the basic structure, and denote the result as $G[VS - a_g]$. From connected basic structures in G, consecutive amino acids can be inferred. Here, "connected basic structures" means that for any basic structure $G[VS - a_g]$ belonging to the whole structure, there is at least one other basic structure $G[VS - a_h]$ having vertices overlapping with $G[VS - a_g]$. For example, in Figure 3.2, $G[VS - a_i]$ and $G[VS - a_j]$ are connected basic structures that have v, v_c, v_{loss} and v_{c-loss} as overlapping vertices.

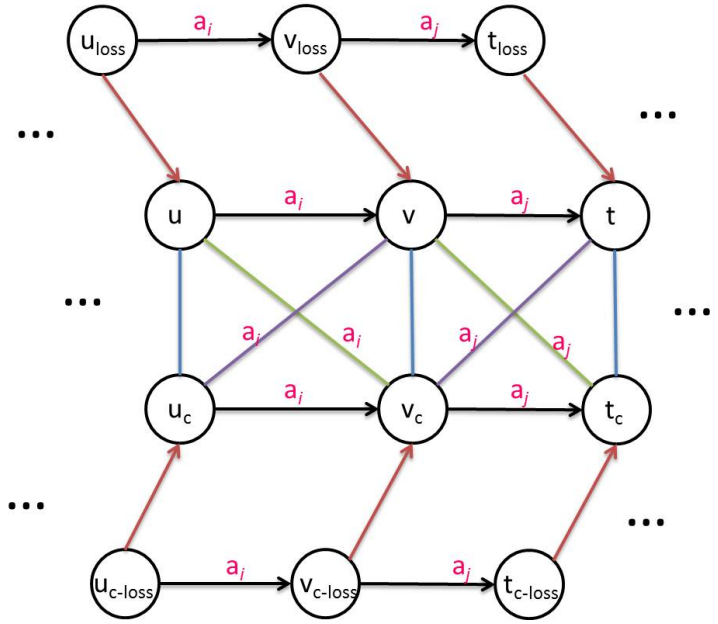


Figure 3.2: An example of a multi-edge graph.

Another graph G_{AA} can be used to represent the relationship of amino acid strings inferred from basic structures, in which the vertices V_{AA} and edges E_{AA} are defined as below.

V_{AA}: Each vertex represents an amino acid acquired from a basic structure of a multi-edge graph.

E_{AA}: Two amino acids are connected by an undirect edge if they are inferred from two connecting basic structures.

With the multi-edge graph G and the basic structures $G[VS]$, the peptide sequencing problem is then transformed into a path-finding problem in a simple graph G_{AA}.

During a MS/MS experiment, spectra representing different kinds of ions, along with their loss of small molecules, are created. Assuming that there are τ types of ions present, the set of these ions is defined as

$$\Delta = \{\delta_1, \delta_2, \ldots, \delta_\tau\}. \tag{3.1}$$

For a peptide $P = a_1 a_2 \ldots a_n$, in order to find its full length sequence using a spectrum graph model, at least one δ ion should be observed at each cleavage site between amino acid a_i and a_{i+1}, $\forall i \in 1, 2, \ldots, n-1$. In the traditional spectrum model, only two kinds of ions, b-ions and y-ions, are considered when finding the longest paths. Here, 8 more types of ions are considered in the multi-edge graph model including loss of H_2O, NH_3, CO- and NH- groups of both b-ions and y-ions. Therefore, the multi-edge graph model has a higher chance of interpreting the full length peptide sequences than the traditional model.

3.2.2 Integration of peptide tags and AAC information

With more types of ions considered, the computational time and the possibility of false positives could be increased. As a solution to this problem, more information from MS/MS spectra can be integrated into the

model. Here, peptide tags and AAC information are incorporated into the proposed model.

Since peptides are long strings of amino acids, a straight-forward idea to make the sequencing problem easier would be cutting the long strings into shorter, more easily interpreted substrings, and then solving the sequencing problem on these shorter strings. Here, NovoHCD uses peptide tags to separate the whole peptide sequences into three parts, namely prefix, tag, and suffix, and then uses the multi-edge graph model on both prefix and suffix with AAC information to limit the number of edges considered in the multi-edge graph model.

Currently, the most widely used peptide tags are 3-tags (tags consisting of three amino acid residues) [22]. Since the major focus of this paper is multi-edge graph based peptide sequencing and not tag finding, and there are already many tag-finding algorithms available, a suitable 3-tag generating algorithm is chosen to output all tags needed in the proposed method. An effective method named DirecTag [22] is used. For the AAC information, an in-house database was built containing all theoretically possible AACs for any given mass of no more than 3000Da, which should be able to output the AACs for almost all peptides encountered in a peptide sequencing problem.

The detailed procedure of inferring peptide sequences from the multi-edge graph model with AAC information and a given peptide tag t is summarized in Algorithms 1 and 2. The following notions are used in the algorithms. T is the set of all tags, $S = \{(m_i, int_i) \mid i = 1, 2, \ldots, n\}$ is the peak list of an experimental spectrum. $PreSeq$ and $SufSeq$ are the sequence sets of the possible prefix and suffix sequences of the original peptide separated by tag t, respectively. Δm_{pre} and Δm_{suf} represent the mass values of the prefix and suffix separated by t, respectively. $G([V, E])$ is a multi-edge graph representing the various relationships between ions, where each vertex in V is a peak in S. $G[VS]$ is a basic structure of $G([V, E])$. A_{now} is the AAC used in the path finding in the current multi-edge graph $G([V, E])$.

Algorithm 1 Peptide sequencing using multi-edge graph model with integration of AACs and known peptide tags

Input: A tag $t \in T$, experimental spectrum $S = \{(m_i, int_i) \mid i = 1, 2, \ldots, n\}$, Δm_{pre} and Δm_{suf}.

Initialize: $G([V, E])^0 \Leftarrow [Ion, \emptyset]$, $Candidate \Leftarrow \emptyset$.

Calculate AACs of Δm_{pre} and Δm_{suf}, and put into sets AAC^{pre} and AAC^{suf}

Calculate the m/z values of first and last ions of t, denoted as $m(it)_s$ and $m(it)_e$

Invoke Algorithm 2 with input AAC^{pre} and $m(it)_s$, yielding $PreSeq$

Invoke Algorithm 2 with input AAC^{suf} and $m(it)_e$, yielding $SufSeq$

for all $seq_p \in PreSeq$ **do**

  for all $seq_s \in SufSeq$ **do**

    Combine seq_p and t and seq_s, and put it into $Candidate$

  end for

end for

Output: $Candidate$.

Algorithm 2 Partial peptide sequence generation based on multi-edge graph model and given AACs

Input: Set of AACs A, m/z values of ion $m(it)$.

Initialize: Set of paths $Seq \Leftarrow \emptyset$.

while $A \neq \emptyset$ **do**

Put first AAC in A into A_{now}

$G([V,E]) \Leftarrow G([V,E])^0$, $Vnext \Leftarrow m(it)$

while $Vnext \neq \emptyset$ **do**

Put first vertex in $Vnext$ into u

for all $m_i \in V$ **do**

Check if any edges of types e_{uv}^1 to e_{uv}^6 can be drawn when using A_{now} between u and m_i, and let e be the number of such edges

Add the e edges to E

Add the vertices connected with u into $Vnext$

end for

end while

Find basic structures $G[VS]$ in $G([V,E])$ and build induced graph G_{AA}

Infer paths from G_{AA} qualified by A_{now}

Add qualified paths into Seq

Delete A_{now} from A

end while

Output: Set of paths Seq.

Table 3.1: Description of mass-based features used in tag ranking

Feature of v	number of ions u in a spectrum where
f_1	u is the complementary ion of v.
f_2	u and v are the +1 and +2 charges of the same partial peptide.
f_3	u has a mass difference of a single amino acid to v.
f_4	u has a mass difference of a loss of a small molecule from v.
f_5	u and v are isotopic pairs.
f_6	mass sum of u and v is close to parent mass with one amino acid loss or overlap.

3.2.3 Ranking of peptide tags

The peptide tag finding algorithm employed by DirecTag contains a criterion for ranking the output tags. However, under scrutiny this ranking criterion was determined to be not very effective, thus limiting the performance of NovoHCD. Therefore, a new peptide tag ranking method is developed here in order to pick the best tags more effectively.

The proposed ranking method borrows ideas from the multi-edge graph model when choosing mass-based features, and also the idea of "local maximum" from other research conducted by our group [88] when applying intensity information to feature calculation. This method first calculates the rank of each peak in a spectrum, and then combines ranks of all four peaks forming a tag to be the rank of the tag. This ranking method utilizes information not considered by DirecTag, and thus is expected to achieve better results. In this method, six mass-based features, denoted as $F = [f_1, \ldots, f_6]$, are first generated. These features are all numbers of ions satisfying the criteria listed in Table 3.1. Calculation of a final score for each ion v, denoted as $F(v)$, is defined as the linear combination of all features,

$$F(v) = \sum_{s=1}^{6} \sum_{u \in E \setminus v} f_s(u, v) \cdot w_s, \tag{3.2}$$

where w_s is the weight of f_s, and E is the edge set of the graph G. The weight vector $W = [w_1, \ldots, w_6] = [1, 0.5, 1, 0.2, 0.5, 1]$ is defined similarly to the weight assignment method used in [88].

It has been shown that a simple threshold is not effective in differentiating signal ions from noisy ions because the ions' intensities in a spectrum tend to be larger in the middle of a m/z range than at the two ends [88]. Rather, it is more reasonable to assume that the noises in a narrow m/z range are described by a single simple distribution (for example, a normal distribution) and that the signal ions tend to be the local maxima [88]. In this circumstance, we define *local maximum* and *order* in the use the intensity of each ion v as follows.

Local maximum: Vertex v is called a "local maximum" if its intensity is larger than the two peaks beside it.

Order: The order of vertex v, denoted as o, is the iteration on which it is picked as a local maximum in the following procedure. $o = 1$ means the peak is a local maximum of the original spectrum. After deleting these $o = 1$ peaks, the local maximum peaks from the remainder of the spectrum will be assigned orders $o = 2$. This process repeats iteratively until all peaks have been assigned with an order.

Finally, in the ranking process, an ion v with order o can be represented as a vector $R(o, F(v))$. We order R first according to its initial element o in an increasing order, and then F in a decreasing order if two o values are the same. After this ranking process, each v has its rank, denoted as r_v. If a tag T consists of four ions v_i, v_j, v_k, and v_t from the original experimental spectrum, the rank of T, denoted as r_T, is calculated as

$$r_T = r_{v_i} + r_{v_j} + r_{v_k} + r_{v_t}. \tag{3.3}$$

The smaller the value of r_T, the greater the chance that tag T is selected in the proposed method.

3.2.4 Candidate peptides scoring scheme

When all candidate peptides have been generated, the last step of NovoHCD is to rank these candidates and determine the most likely correct candidates. Parent peptide mass m_P is a widely-used feature to filter out incorrect candidates or bound the search space. Therefore, the mass difference between candidates and precursor ions (given in the spectrum), denoted as Δm_P, is used as a key feature in the ranking. In addition, IMs can be observed in HCD spectra, and possible amino acids can be inferred from these ions. Each IM corresponds to a single amino acid. All amino acids inferred from IMs observed in a spectrum are defined as AA_{IM}, and the number of amino acids in AA_{IM} that are included in the candidate sequence is defined as N_{IM}. Therefore, for a given spectrum, each candidate peptide generated by NovoHCD can be represented as a vector $CP(N_{IM}, \Delta m_P)$. An approach similar to the one in the proposed tag ranking method is then used. It first sorts CP according to its initial element N_{IM} in a decreasing order, and then arranges Δm_P in an increasing order. After this process, each candidate and its final ranking is output.

3.3 Experiments and Results

Five MS/MS datasets were used to test the performance of the proposed method. The performance was then compared with that of another *de novo* peptide sequencing algorithm for HCD spectra named pNovo [72]. The detailed experimental process and comparison results are presented below.

3.3.1 Datasets

Five datasets were used to investigate the performance of our new method and pNovo. The first dataset, SwedHCD contains +2 charged high-energy (HCD) MS/MS spectra of unique peptides [97]. The other

four datasets – SCX_decon, SCX_nodecon, 60min_analyses, 300min_analyses – are from the same research paper [98]. The original datasets contain various fragmentation MS/MS spectra including CID, HCD, and ETD spectra. The HCD spectra were selected from them for the experiments here. The number of spectra from the above five datasets used in the analysis were 10878, 1952, 2557, 123 and 154, respectively. Spectra from the latter four datasets have charges of +2, +3, and +4.

To determine correctness of the peptides produced by the two programs for each input spectrum, a "gold standard" was necessary. Every spectrum in the five datasets came with a correct spectrum. To give a more comprehensive comparison, sequences from Mascot were also used as correct peptide sequences. Mascot [99] is a widely-used program for protein identification based on database search. All results produced by Mascot were trimmed to a $< 5\%$ false discovery rate.

3.3.2 Peptide tags ranking performance

The performance of the peptide tag ranking methods was first investigated: our proposed method versus DirecTag. The SwedHCD dataset was used for this. For each spectrum, DirecTag produced 50 tags with their rankings. To evaluate the two methods, we compared the number of spectra having at least one correct tag under the condition of the same number of tags selected. The comparison when selecting at most 30 tags per spectrum is shown in Figure 3.3. The y-axis in Figure 3.3 is the number of spectra having at least one correct tag.

Sequences from Mascot were used as the correct sequences when generating the results in Figure 3.3. Using the correct sequences as supplied by the databases gave similar results (data not show). Figure 3.3 shows that given the same number of (top-ranked) tags selected, more spectra contain at least one correct peptide tag using NovoHCD. For example, when selecting 30 tags for both methods, there were 9,384 spectra having at least one correct tag by using NovoHCD, while the number was 9,176 by using DirecTag's ranking criterion. Similarly, if we set the number of peptides having at least one correct tag to be the same, NovoHCD needed fewer tags output. For example, when setting the peptide number to be 9,176, NovoHCD needed no more than 21 tags per spectrum, while the DirecTag needed 30 tags per spectrum.

The proposed ranking method makes a significant contribution to the whole proposed multi-edge graph based method. Using this ranking method, higher peptide sequencing accuracy can be achieved with fewer tags, which corresponds to fewer multi-edge graphs calculated and fewer candidate peptides generated. Thus, the computational costs and number of false positive candidate peptides can be reduced by using the proposed ranking method.

3.3.3 *De novo* peptide sequencing performance

All five datasets were used to investigate the performance of the proposed method and pNovo. For each spectrum, the top three candidates output by NovoHCD and pNovo were selected for evaluation. If any one of the three candidates was correct, we said that this spectrum achieved a full length accuracy for the given

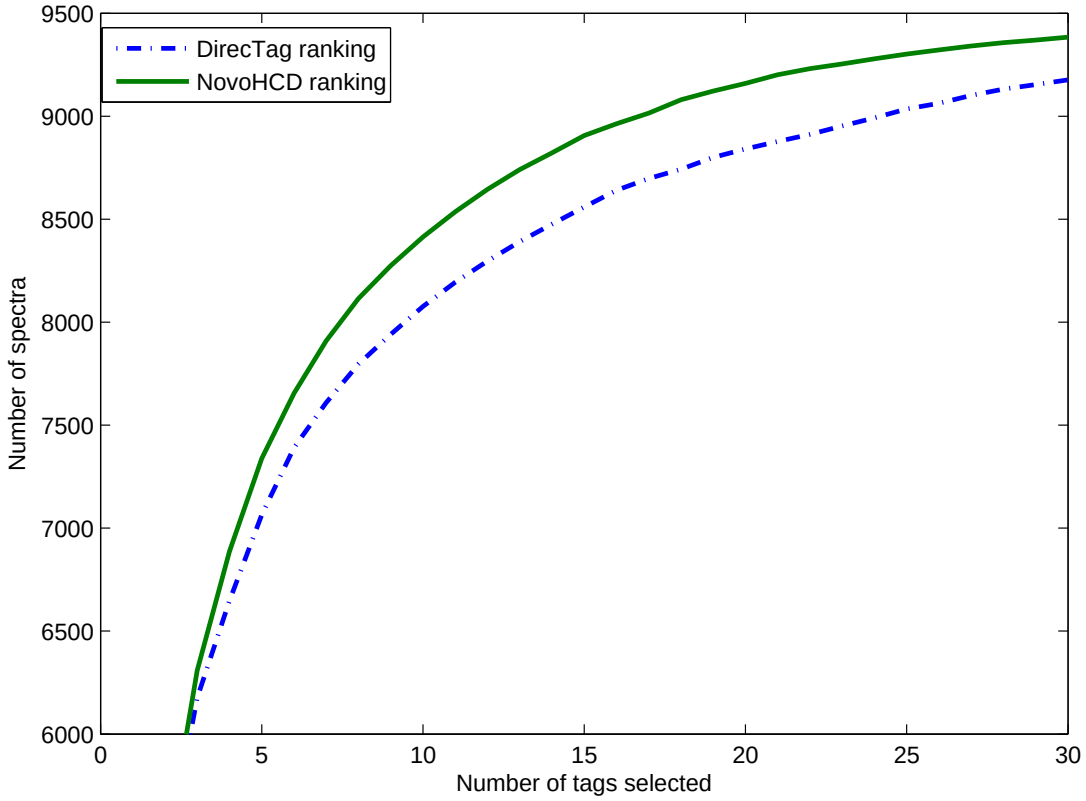


Figure 3.3: Performance comparison of NovoHCD's ranking method (solid green curve) and DirecTag ranking (dashed blue curve).

method. The results for the full length peptide sequencing accuracy comparison are presented in Tables 3.2 and 3.3 with the use of Mascot and datasets providing correct sequences, respectively.

From Tables 3.2 and 3.3 one can see that for all five datasets, NovoHCD achieves higher full length accuracy than pNovo, and the improvement ranges from 13% to 21% on different datasets. On SwedHCD dataset, both methods achieved highest full length accuracy. This could be because this dataset only contains +2 charged peptides, which makes the spectra less complex and easier to process. In addition, both methods achieve lowest accuracy on the 300min_analyses dataset. This might be due to the low quality of the spectra, but further investigation is needed to get a reliable explanation. On the 60min_analyses dataset, the two methods have the largest accuracy difference, which might be because that +3 and +4 charged spectra make the sequencing problem more complicated. In this case, NovoHCD considers the +2 charged ions, which occur more commonly in +3 and +4 spectra, thus making its sequencing result better than pNovo's. Other features from +3 and +4-charged peptides will be further studied to improve our method in future study.

Furthermore, we considered the relationship between the number of correctly identified peptides and peptide length. Figures 3.4 and 3.5 summarize the results from comparing the outputs of NovoHCD and pNovo on the two largest datasets of the experiments, SwedHCD and SCX_nodecon, respectively.

Table 3.2: Full length peptide sequencing accuracy comparison among different datasets with Mascot results as correct sequences

Dataset	pNovo	NovoHCD
SwedHCD	80.93%	**94.50%**
SCX_decon	65.89%	**81.46%**
SCX_nodecon	77.90%	**88.46%**
60min_analyses	60.87%	**81.52%**
300min_analyses	53.44%	**67.17%**

Table 3.3: Full length peptide sequencing accuracy comparison among different datasets with sequences from datasets as correct sequences

Dataset	pNovo	NovoHCD
SwedHCD	80.95%	**94.54%**
SCX_decon	63.11%	**80.58%**
SCX_nodecon	73.99%	**85.14%**
60min_analyses	61.79%	**81.30%**
300min_analyses	50.65%	**67.53%**

Figures 3.4 and 3.5 show that NovoHCD identified more peptides at every peptide length, as compared to pNovo. The most peptides identified have lengths 8 to 10 with the SwedHCD dataset, and 13 to 16 with the SCX_nodecon dataset. In addition, the SCX_nodecon dataset has longer peptides than the SwedHCD dataset, and an increase of peptide length corresponds to fewer correct peptides inferred by both methods. One can see that for length larger than 25 in the SCX_nodecon dataset, almost no peptides were correctly identified by either method. Therefore, long peptide sequencing is still a challenging problem in *de novo* peptide sequencing. NovoHCD, which breaks whole peptides into three shorter parts and then infers their sequences, partly solves this problem, thus achieving better full length sequencing results when compared to pNovo.

Finally, we examine the computational time of NovoHCD. The algorithm was written using Matlab (2010b) and the program was run on a PC with a 3.07 GHz quad-core CPU (Windows 7 operating system). The results for different datasets are shown in Table 4.8. All run times (CPU times) are given in seconds.

From Table 4.8 one can see that the computational time varies from dataset to dataset, but all in an acceptable time range. The time difference may be due to varied properties of the datasets, such as the number of peaks per spectrum and the quality of the spectra. Reducing computational time will require

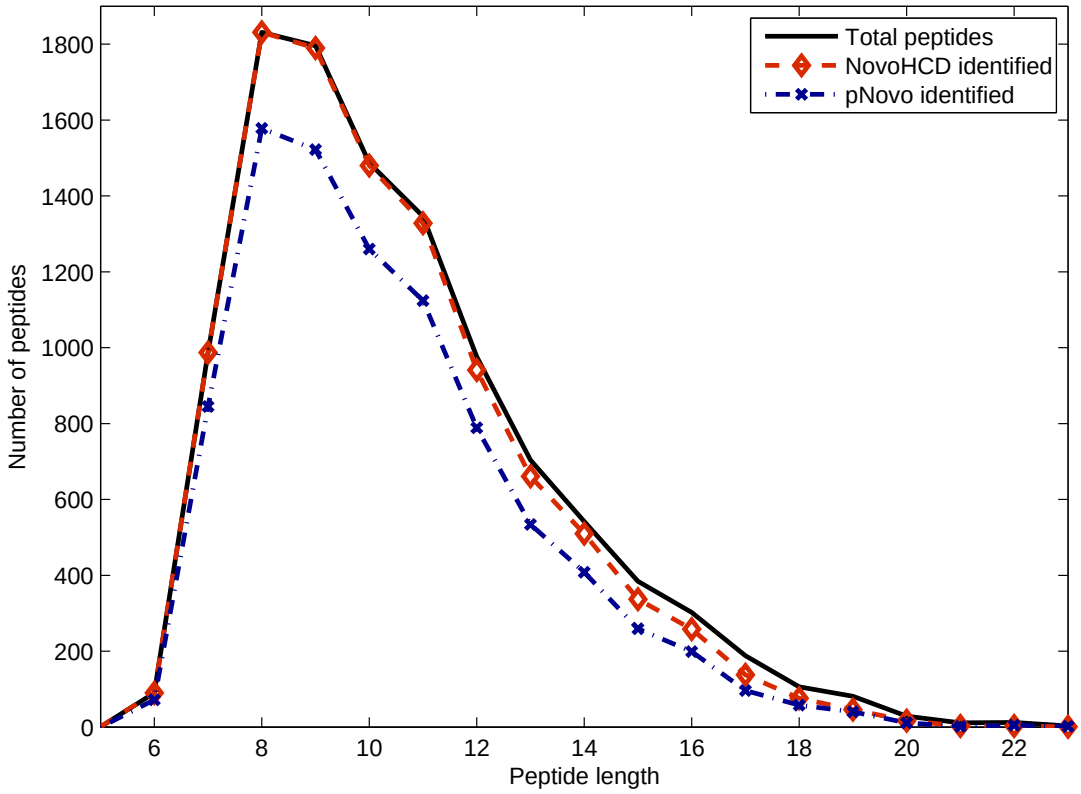


Figure 3.4: Relationship between the number of correctly identified peptides and peptide length for NovoHCD and pNovo using the SwedHCD dataset.

investigation to find the most time-consuming spectra or algorithm steps. Such investigation will be part of our future study of this *de novo* peptide sequencing approach.

3.4 Conclusions and future work

In this paper, a new solution to the *de novo* peptide sequencing problem for HCD spectra, NovoHCD, has been proposed. It is based on multi-edge graphs with integration of AAC and peptide tags. NovoHCD first uses peptide tags to separate a whole peptide sequence into three parts: prefix, tag, and suffix. It then builds multi-edge graph models on prefix and suffix information separately for sequence interpretation, and AAC information is used in limiting edges of the graph. It finally combines these three parts to generate complete sequences of candidate peptides. Immonium ions observed particularly in HCD spectra are used in the candidate peptide ranking of NovoHCD. Five HCD spectra datasets were used to investigate the performance of NovoHCD and compare it with another successful *de novo* peptide sequencing method, pNovo. Experimental results showed that the overall accuracy increases from 13% to 21% compared to pNovo. In addition, NovoHCD also includes a better peptide tag ranking algorithm as compared to another software

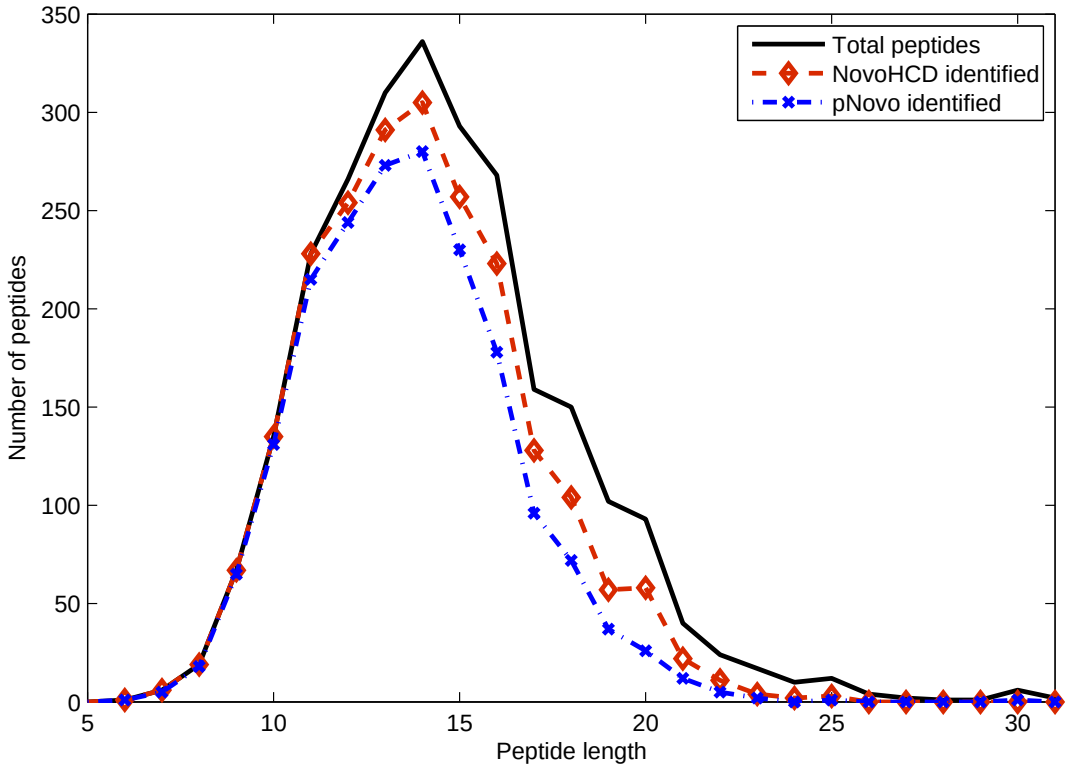


Figure 3.5: Relationship between the number of correctly identified peptides and peptide length for NovoHCD and pNovo using the SCX_nodecon dataset.

called DirecTag, which makes the computation faster and more accurate.

We summarize the major contributions of NovoHCD in the following. First, it replaces the traditional spectrum graph with a multi-edge graph, thus including more relationships between two peaks in a spectrum. Secondly, it separates the whole peptide sequences into three parts, and solves each sequencing subproblem separately. It is straightforward to think of solving a complex problem by cutting it into smaller and simpler subproblems, but an MS/MS spectrum itself cannot be cut into pieces and interpreted since the ion signals are intermixed. Separating a peptide sequence (used in NovoHCD) instead of its spectrum is a more suitable approach. Thirdly, AAC information is effectively incorporated into the peptide sequencing process. AACs have been used before in peptide sequencing, but the AACs derived from the parent peptide mass may result in too many combinations to calculate. However, with the separation strategy above, the number of AACs can be limited to an acceptable number. Therefore, AACs can be used in peptide sequencing with acceptable calculation cost in NovoHCD. Last but not least, NovoHCD has been applied on a new kind of spectra (HCD), and incorporates unique features of this kind of spectrum (immonium ions) into the sequencing method.

In future, we will focus on improving the accuracy of the proposed method by further analyzing the wrongly identified spectra, expanding the evaluation of the new method to more MS/MS datasets, and extending the method to other types of MS/MS spectra, e.g. ETD spectra. In addition, since a popular

Table 3.4: Running time comparison on different datasets using NovoHCD

Dataset	Number of spectra	Total time (in seconds)	Time per spectrum (in seconds)
SwedHCD	10878	3378.32	0.31
SCX_decon	1952	3857.83	1.97
SCX_nodecon	2557	13238.77	5.17
60min_analyses	123	204.87	1.66
300min_analyses	154	313.40	2.03

approach in peptide sequencing is pairing up different types of spectra from the same peptide to obtain more information, we are planning to modify the model and apply it to the multiple spectra sequencing problem.

Acknowledgment

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

Addendum

Amino acid combinations (AACs) are introduced and used in the proposed method NovoHCD. In order to get all possible AACs of a certain mass value, we used computer software to generate all theoretically possible AACs up to length k (the value of k is explained in the following content) and calculate the mass value of each AAC. The maximum value of the AACs used is 3000Da. Since the minimum mass of the 20 standard amino acid is around 57.05Da (glycine), the maximum length $k = 53$. The masses of the AACs having the same integer part are stored into the same folded labelled with that integer number. After this process, when all the AACs of a certain mass value are needed, we can first check and find the correct folder, and then search for all needed AACs.

Biological meanings of of 5 types of edges defined in the proposed NovoHCD are listed below.

Edge Type	Meaning
Type 1	u and v have a mass difference of some amino acid.
Type 2	u and v represent a complementary ion pair.
Type 3	u and v represent a complementary ion pair with one amino acid overlap.
Type 4	u and v represent a complementary ion pair with one amino acid gap.
Type 5	u and v are the same fragment ions but with loss of a small molecule.

Erratum

In the second paragraph of Subsection 3.2.2, “data not show” should be changed to “data not shown”.

Chapter 4

NovoExD: *De novo* peptide sequencing for ETD/ECD spectra

Published as: Yan Yan, Anthony J. Kusalik and Fang-Xiang Wu. "NovoExD: *De novo* peptide sequencing for ETD/ECD spectra," IEEE/ACM Transactions on Computational Biology and Bioinformatics, in press, online available at `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7005483`.

In the previous chapter, a new *de novo* peptide sequencing method for HCD spectra, NovoHCD, is presented. NovoHCD applies a new spectrum graph model with multiple types of edges, and integrates other information like AACs and peptide tags to help with the sequencing. Apart from HCD spectra, there are other types of newly available spectra that have different properties than HCD spectra. Among them, ECD/ETD are widely used. ECD/ETD spectra have variants of c-ions and z-ions as dominant fragment ions, and usually produce high quality MS/MS spectra for multiple charged ($\geq$+3) peptides. These features make ECD/ETD spectra popular for MS/MS based peptide sequencing since they are quite different from the traditional CID spectra and the previously studied HCD spectra.

Based on the review in Chapter 2, current studies in ECD/ETD spectra focus on property analysis of these spectra and sequencing using them and CID spectra together. Typically, the methods designed for the use of CID with ECD/ETD fail to fully consider the features of ECD/ETD spectra. These methods treat ECD/ETD spectra as supplementary to CID spectra and focus on getting the information that CID spectra are missing from the accompanying ECD/ETD spectra. At this time, less attention has been paid to *de novo* sequencing methods for ECD/ETD spectra solely. Facing this situation, a suitable method using ECD/ETD spectra alone is noteworthy.

This chapter presents a new *de novo* peptide sequencing for ETD/ECD spectra named NovoExD. The success of NovoHCD presented in the previous chapter shows the effectiveness of the new designed graph model. It is shown that with suitable improvement and modification of the model and specific feature consideration for ECD/ETD spectra, the new method for ECD/ETD spectra achieves better performance than other existing methods.

The proposed NovoExD in this chapter modifies the graph model in the use of peptide tags and adds corresponding preprocessing steps. To be specific about tag usage, NovoExD changes the single tag usage in NovoHCD to multiple peptide tags. These tags work together to separate a whole peptide sequence that needs

to be identified into smaller pieces. The method infers partial sequences separately and assemble them back together. Computation complexity of NovoExD and the number of possible AACs can be reduced with the use of multiple tags. When considering unique features in ECD/ETD spectra, a charge determination step for fragment ions is designed since a lot of ions in ECD/ETD spectra have multiple charges. Experimental results on three MS/MS spectral datasets show that NovoExD outperforms another competing method in terms of average full length peptide sequencing accuracy.

Abstract

De novo peptide sequencing using tandem mass spectrometry (MS/MS) data has become a major computational method for sequence identification in recent years. With the development of new instruments and technology, novel computational methods have emerged with enhanced performance. However, there are only a few methods focusing on ECD/ETD spectra, which mainly contain variants of c-ions and z-ions. Here, a *de novo* sequencing method for ECD/ETD spectra, NovoExD, is presented. NovoExD applies a new form of spectrum graph with multiple edge types (called a GMET), considers multiple peptide tags, and integrates amino acid combination (AAC) and fragment ion charge information. Its performance is compared with another successful *de novo* sequencing method, pNovo+, which has an option for ECD/ETD spectra. Experiments conducted on three different datasets show that the average full length peptide identification accuracy of NovoExD is as high as 88.70%, and that NovoExD's average accuracy is more than 20% greater on all datasets than that of pNovo+.

4.1 Introduction

Tandem mass spectrometry (MS/MS) is used as a major tool for peptide identification in current proteomics studies. In a typical MS/MS experiment, protein mixtures are first digested into suitably sized peptides, and then the peptides are ionized via an ionization process. After that, selected peptides are further broken into fragment ions, and their tandem mass spectra (MS/MS spectra) are collected [8]. MS/MS spectra usually contain two kinds of information for each ion detected, its mass-to-charge (m/z) value and intensity.

In a typical MS/MS experiment, peptide ions are broken into various kinds of fragment ions. There are six kinds of commonly observed ions, namely a-, b-, c-, x-, y-, and z-ions. Different fragmentation techniques in MS/MS yield different dominant types of fragment ions. Collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD) yield b-ions and y-ions as dominating ions. Electron capture dissociation (ECD) and electron transfer dissociation (ETD) preferentially produce variants of c-ions and z-ions, and occasionally a-ions [31–33]. ETD [37] is a modification of the ECD technique [38] that was designed for dissociation of multiply protonated peptide ions in MS/MS. In this paper, we use ExD to represent ECD and ETD spectra as a whole.

CID was the most commonly used fragmentation technique when researchers started using mass spectrometry for peptide sequencing. With the development of new techniques and instruments in recent years, alternative MS/MS spectra with new features have appeared. Among these, ExD spectra have drawn increasing attention because of their unique features. Specifically, ExD produces high quality MS/MS spectra for multi-charged peptides and has no strong cleavage preferences. It utilizes a lower energy pathway than CID and HCD, thus preserving labile post-translational modifications (PTMs) [39–41]. All these features yield spectra containing useful information, and hence they have the potential to give satisfying peptide sequencing

performance.

Currently, there are three main kinds of methods used for peptide sequencing from MS/MS data: database searching, peptide tagging and *de novo* sequencing. In database searching, theoretical spectra are computed from an existing protein database and peptides are identified by matching the theoretical spectra to experimental spectra [46]. The major disadvantage of database searching is that it cannot identify new or unknown peptides. Peptide tagging [21, 22] usually produces partial sequences, often called tags, from an MS/MS spectrum, and then uses these tags to search against a protein database or to help with *de novo* sequencing. The use of tags can dramatically reduce the search space and time needed, and has the potential to improve *de novo* peptide sequencing. *De novo* sequencing automatically interprets spectra using the masses of amino acids. It can identify new proteins, proteins resulting from mutations, proteins with unexpected modifications and so on. With the recent development of high mass-accuracy MS/MS and alternative fragmentation techniques, *de novo* sequencing has shown promising developments [48]. Therefore, this study focuses on *de novo* peptide sequencing.

ExD is a new technology that has properties different from CID and HCD. The recent studies of MS/MS spectra produced by ExD have focused on the characteristics of the spectra [32, 33, 39], how various PTMs can be identified from the spectra [40], and the performance of such spectra [17, 74]. When used for peptide sequencing, ExD spectra are often paired with CID (or HCD) spectra because of the availability of the complementary information from these spectra [31, 75, 77, 79]. Some of the methods that use paired spectra for sequencing also come with an option to use ExD spectra alone; for example, the one introduced in [75]. These methods typically consider only a subset of the features in ExD spectra because their focus is paired spectra sequencing. At this time, less attention has been paid to *de novo* sequencing methods using ExD spectra alone, especially when compared to the ones developed for CID or HCD spectra alone. Considering the unique features of ExD spectra and the shortage of corresponding algorithms, a suitable method designed for ExD spectra is useful and noteworthy.

In this paper, we present a modified spectrum graph with multiple edge types, denoted as GMET, derived from [100] to model ExD spectra. Amino acid combinations (AACs), the order-independent amino acid composition information of a peptide, and peptide tags have already been integrated in our previous model [73]. This new model utilizes all previously considered features as well as unique features in ExD spectra. Since it is quite common to acquire ExD spectra from multi-charged peptide samples, multi-charged fragment ions are frequently observed. Wisely considering and utilizing these ions can be a great help in peptide sequencing.

The remainder of the paper is organized as follows: Section 2 presents the overall design, the GMET model, and schemes for charge determination of fragment ions, peptide tags and amino acid usage, and the candidate peptide ranking. Section 3 presents an evaluation of the method. Shown are the evaluation methodology, the experimental results and performance analysis. Finally Section 4 concludes this study and gives some directions for future work.

4.2 Methods

In this section, a new *de novo* peptide sequencing algorithm for ExD spectra is proposed. The algorithm considers unique features of ExD spectra including alternate dominant ions and multi-charged fragment ions, and uses a recently presented type of graph, a graph with multiple edge types (GMET) [73]. The graph is constructed by considering multiple peptide tags and fragment ion charge information.

The whole method is summarized in Figure 4.1. To start, the algorithm first searches the low mass region of a spectrum, typically less than 200Da (Dalton, the unified atomic mass unit), to find any consecutive, same-type ions from the first two cleavage points; to be specific, ion pairs $\{c_1, c_2\}$, $\{z_1^{\cdot}, z_2^{\cdot}\}$, $\{c-1_1^{\cdot}, c-1_2^{\cdot}\}$, and $\{(z+1)_1, (z+1)_2\}$. In this notion for ions, "+1" and "-1" represent addition or loss of 1Da in mass, respectively; subscript numbers indicate the cleavage positions on a peptide backbone from either N-terminal or C-terminal; and an ion with a dot "$\cdot$" means a radical fragment ion, which is a free radical species carrying a charge. If any pair exists, the two amino acids at the ends of a peptide sequence can be inferred and the sequencing will be conducted on the rest of the peptide sequence. The motivation of this preprocessing step is to reduce the potential sequencing length and make the problem easier to solve. This step is shown in Figure 4.1 in a dashed box indicating that one may not find any ion pairs for some spectra to limit the sequencing length, especially for the spectra with poor fragmentation at the terminal amino acids.

After this preprocessing step, a tag finding and ranking algorithm is applied to find length-3 tags and their associated scores. The tags are sorted in a decreasing order of the tag scores, and placed in a set T. A large score corresponds to high confidence that a tag belongs to the peptide that generated the specific spectrum.

The algorithm then uses the first peptide tag in T to separate prospective whole sequences into smaller pieces. If there are no unused tags in T, the whole algorithm stops after a candidate peptide ranking step; else, the first tag t in T is selected and deleted from T. Then the two mass regions separated by t are calculated, stored in set $Part$, and compared to a predefined threshold $Thres$. If any of the mass regions is over $Thres$, further separation of the region is needed using another suitable tag $t^{'}$ in T; else, all amino acid compositions of all regions in $Part$ are calculated. Here, any length-3 tag consists of four consecutive ions. The mass values of the two end ions of $t^{'}$ should be in the interval of the mass region (over $Thres$) in order to separate it.

After that, the algorithm calculates amino acid compositions (AACs) of each region in $Part$ and builds GMETs on these regions to find partial peptide sequences. Finally, all suitable parts are assembled together to form the final peptide candidates, and a ranking scheme is applied to select and output the best ones. These steps in the algorithm and the various variables appearing in Figure 4.1 are explained in detail in the following subsections.

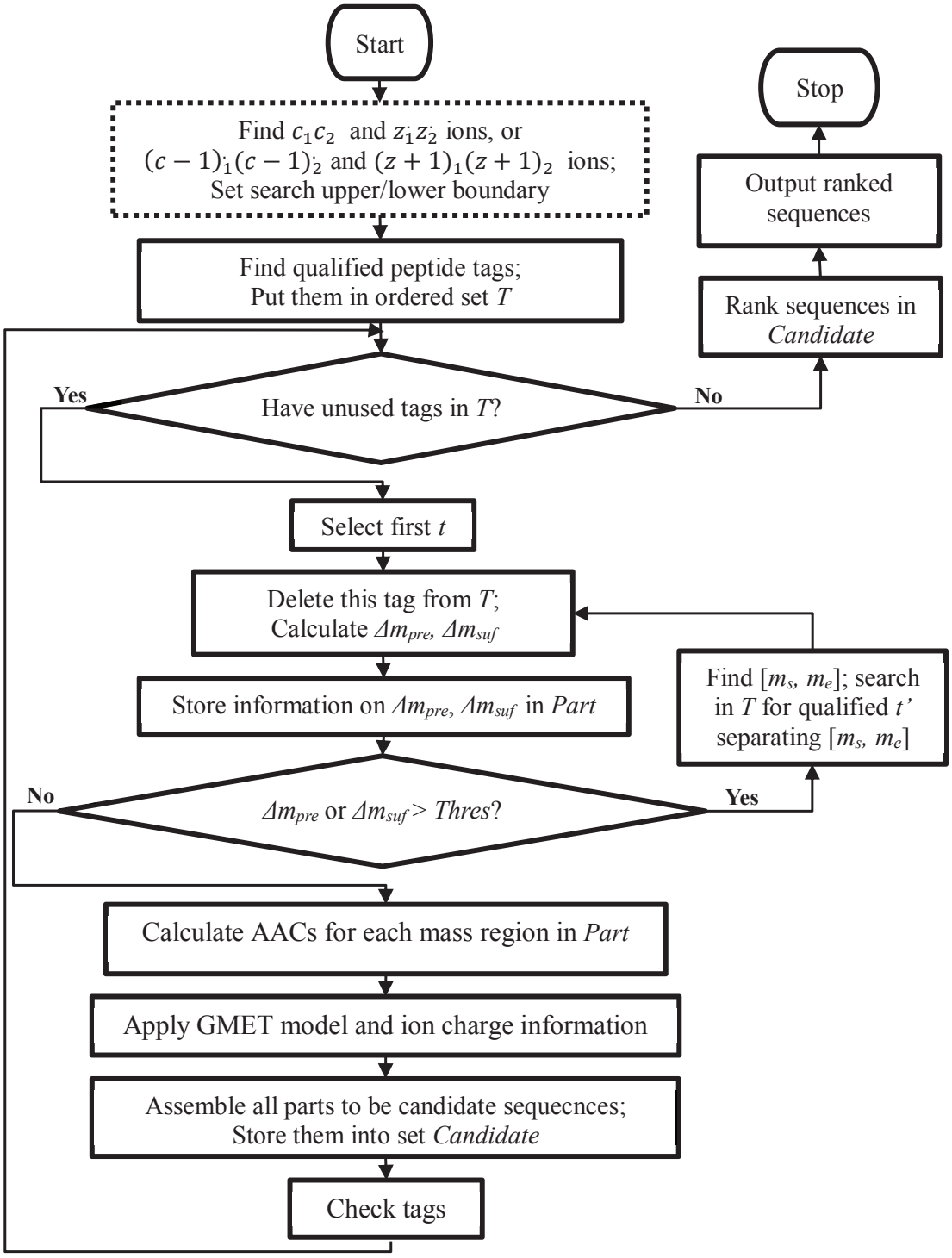


Figure 4.1: Method flow chart

Table 4.1: Ion types considered in ExD spectra

Ion Type	Mass calculation from residues	Mass calculation from other ions
c	$\sum(residue\ mass) + 18.0344$	$b_m + m_{NH_3}$
$c-1^{\cdot}$	$\sum(residue\ mass) + 17.0265$	$b_m + m_{NH_3} - m_H$
$z^{\cdot}$	$\sum(residue\ mass) + 3.0156$	$y_m - m_{NH_3} + m_H$
$z+1$	$\sum(residue\ mass) + 4.0156$	$y_m - m_{NH_3} + 2m_H$
w	$\sum(previous\ residue\ mass) + 73.0290$	$x_{previous} + m_{CO}$
b	$\sum(residue\ mass) + 1.0078$	b_m
y	$\sum(residue\ mass) + 19.0814$	y_m

4.2.1 Basic ion types considered in ExD spectra

Since ExD spectra have dominant ion types different from CID and HCD spectra, we first investigate the ion types considered for the proposed method. After study of the literature about the frequency of different fragment ions [32,33,37,39,41,87], the ions listed in Table 5.2 were selected based on their frequency observed in MS/MS spectra. Here, m_{NH_3}, m_H, m_{CO} denote the mass of NH_3, H, and the CO-group, respectively. $\sum(residue\ mass)$ is the mass sum of all amino acids from the end amino acid of a peptide sequence to the amino acid at the current cleavage point. Similarly, $\sum(previous\ residue\ mass)$ is the mass sum of all amino acids from the end amino acid of a peptide sequence to one amino acid previous to the current cleavage point. b_m and y_m are the masses of the b-ion and y-ion, respectively, at the current cleavage point. $x_{previous}$ is the mass of the x-ion at the cleavage point one position previous to the current one. In addition, the complementary ion relationships in Equations (1)-(3) hold for the ions in Table 5.2.

$$b_i + y_{N-i} = m_p + 2m_H, \tag{4.1}$$

$$c_i + z^{\cdot}_{N-i} = m_p + 3m_H, \tag{4.2}$$

$$c-1^{\cdot}_i + (z+1)_{N-i} = m_p + 3m_H, \tag{4.3}$$

where m_p is the mass of parent peptide P, N is peptide length, and $i \in \{1, 2, \ldots, N\}$. Δ_i is the ion mass of the i^{th} Δ-ion, where $\Delta \in \{c, c-1^{\cdot}, z^{\cdot}, z+1, b, y\}$.

4.2.2 Charge determination of multi-charged fragment ions

Since ExD spectra resulting from multi-charged peptides are quite common, fragment ions in spectra are often in multiple charge states. Fragment ions having different charges mixed together in a spectrum may cause interpretation problems since the mass difference calculation, a key measurement used in *de novo*

peptide sequencing, is affected. Therefore, determination of fragment ion charges is quite useful in peptide sequencing.

There are already a lot of methods developed for MS/MS spectrum charge deamination, and many of them focus on the charge determination of precursor ions and utilize sophisticated machine learning strategies [92, 93, 101, 102]. In this paper, we use a straightforward method to determine charges of all fragment ions in a spectrum in order to reduce the complexity and reduce the computational time of the proposed NovoExD.

Let the charge of a spectrum S from a peptide P be a non-negative integer n. The charge of a fragment ion i, denoted ξ_i, in this spectrum is then a member of $\Xi = \{+1, +2, \ldots, n-1\}$. Ions having charge n are not considered since the majority of ions in S do not have such a charge value. To determine ξ_i without knowing the peptide sequence, ion relationships are needed. Here, two different possible relationships are considered. One is that the two ions are the same fragment ion with different charges, e.g. $+1$ and $+2$ of the same c-ion. The other is that the two ions are complementary but with different charges, e.g. a $+1$ c-ion and the complementary $+2$ $z^{\cdot}$-ion. These relationships are used to determine (part of) the ion charges in a spectrum.

In a spectrum S with charge n, suppose fragment ion i has charge $\xi_i \in \Xi$. Therefore, we have $n-1$ possibilities for the charge of an ion $i \in S$. Then for S, we build $n-1$ scenarios assuming all ions in S are in charge state $\xi_i \in \Xi$, and calculate the associated $n-1$ spectra having all ions with charge state $+1$, denoted as $S_{\xi_i to 1}$. The set of these $n-1$ spectra is denoted $SA = \{S_{\xi_i to 1} \mid \xi_i \in \Xi\}$. Let j be an ion in $S_{\xi_i to 1}$. Then $\forall j \in S_{\xi_i to 1}$, we have $\xi_j = 1$, which means that all ions in $S_{\xi_i to 1}$ have charge values $+1$. The ion charges can be inferred by considering the two relationships described in the previous paragraph. If two ions $p \in S_{\xi_p to 1}$ and $q \in S_{\xi_q to 1}$ satisfy either of the relationships, where $\xi_p, \xi_q \in \Xi$, then the associated ions of p and q in the original spectrum S, denoted as i_p and i_q, are in charge state ξ_p and ξ_q, respectively. The detailed steps of charge determination of multi-charged fragment ions are shown in the following. Here, the m/z value of an ion $i \in S$ is denoted as $(m/z)_i$.

1. For a spectrum S with charge n, calculate spectrum $S_{\xi_i to 1}$ consisting of only charge $+1$ ions. $\forall j \in S_{\xi_i to 1}$, we have $(m/z)_j = \xi_i * (m/z)_i - (\xi_i - 1)$, $\forall i \in S$, and $\forall \xi_i \in \Xi$.

2. For ions $p \in S_{\xi_p to 1}$ and $q \in S_{\xi_q to 1}$, calculate $M_{diff1} = |(m/z)_p - (m/z)_q|$, where $\xi_p \neq \xi_q$. If $M_{diff1} < \delta_1$, then the associated ions of p and q in S, denoted as i_p and i_q, are in charge state ξ_p and ξ_q, respectively. δ_1 is a small valued threshold. This calculation is between two of the $n-1$ spectra in SA indicating the same ion with different charges.

3. For ions $p \in S_{\xi_p to 1}$ and $q \in S_{\xi_q to 1}$, calculate $M_{diff2} = |(m/z)_p + (m/z)_q - m_p - 3m_H|$, where $\xi_p + \xi_q = n$. If $M_{diff2} < \delta_2$, then the associated ions of p and q in S, denoted as i_p and i_q, are in charge state ξ_p and ξ_q, respectively. δ_2 is a small valued threshold. This calculation can be on the same spectrum in SA or two spectra in SA forming a complementary ion pair.

4. Assign all identified ξ_i to i, and set $\xi_j = 1$ for the rest of the ions $j \in S$; output results.

We give a simple example to show how the charges are determined. Assume all m/z values from an experimental spectrum are stored in set $S = \{73, 109, 116, 130, 183, 217, 346, 365\}$. The parent mass is $m_p = 492$, and the spectrum charge is +4. Here, we use integers to simplify the calculation and focus on the principle of the algorithm.

Since the spectrum charge is +4, from the above algorithm three assumptions are made for the spectrum that assume all ions are in charge states +1, +2, and +3, respectively. Then the three associated spectra having all ions with charge state +1 are calculated as:

$$S_{1to1} = S = \{73, 109, 116, \underline{130}, 183, \mathbf{217}, \mathbf{346}, 365\},$$
$$S_{2to1} = \{145, \mathbf{217}, 231, 259, \underline{365}, 433, 691, 729\},$$
$$S_{3to1} = \{\mathbf{217}, 325, \mathbf{346}, 388, 547, 649, 1036, 1093\}.$$

From the above sets, we first find identical elements which indicate the same fragment ion in different charges. Number 217 (in boldface above) occurred 3 times and 346 occurred twice. From the name of the set and the element positions, we infer that 73, 109 and 217 are the same ion in charge states +3, +2 and +1, respectively; and 116 and 346 are the same ion in charge states +3 and +1, respectively. Subsequently, we use complimentary ion relationships to infer ion charges of the ions in S. Since values 130 (underlined above) in S_{1to1} and 365 in S_{2to1} satisfy such a relationship, we infer that 130 is in state +1 and 183 (the ion at the same position of ion 365 in S) is in state +2. Finally, we label the only remaining ion 365 in S with charge +1. After this process, every ion in S has a charge value.

4.2.3 GMET Model and *de novo* sequencing procedure

This GMET model used here is derived from [73]. The new graph type is of the form GMET=(V, E, Ξ), where each peak (corresponding to a fragment ion) in an experimental spectrum is represented as a vertex $v \in V$ and its m/z value is denoted as $(m/z)_v$; each v has a charge value $\xi \in \Xi$, which is determined by the charge determination process described above.

Five different types of edges (see Table 4.2) in E are considered in the GMET, and the detailed calculations are described in the following. Here, m_p denotes the mass of the peptide producing such an experimental spectrum. A denotes the set of 20 amino acids, and $a_i \in A$ is a certain amino acid. a_i is also used to represent its residue mass. m_{loss} is defined to be the mass of some small molecules or groups lost from fragment ions, which typically include H_2O, NH_3, CO- and NH- groups. $\forall u, v \in V$, if m_u and m_v denote their mass values, then we have $m_u = ((m/z)_v * \xi_u) - (\xi_u - 1)$ and $m_v = ((m/z)_v * \xi_v) - (\xi_v - 1)$. θ is a given threshold. When the edge calculation in the GMET involves an amino acid mass, the fitted amino acid labels the edge.

1. Type I edge, e^I_{uv}: A directed edge e^I_{uv} is drawn from u to v when $\mid (m_v - m_u) - a_i \mid \leq \theta$.

2. Type II edge, e^{II}_{uv}: An undirected edge e^{II}_{uv} is drawn between u and v when $\mid m_p + 3m_H - (m_v + m_u) \mid \leq \theta$.

Table 4.2: Edge types in the GMET

Edge	Relationship
Type I	amino acid difference
Type II	u and v represent a complementary ion pair
Type III	u and v represent a complementary ion pair with one amino acid overlap
Type IV	u and v represent a complementary ion pair with one amino acid gap
Type V	loss of a small molecule

3. Type III edge, e_{uv}^{III}: An undirected edge e_{uv}^{III} is drawn between u and v when $\mid (m_v + m_u) - (m_p + 3m_H) - a_i \mid \leq \theta$.

4. Type IV edge, e_{uv}^{IV}: An undirected edge e_{uv}^{IV} is drawn between u and v when $\mid m_p + 3m_H - (m_v + m_u) - a_i \mid \leq \theta$.

5. Type V edge, e_{uv}^{V}: A directed edge e_{uv}^{V} is drawn from u to v when $\mid (m_v - m_u) - m_{loss} \mid \leq \theta$.

In the experiments described in this study, $\theta = 0.01$ Da. However, the threshold can also be set by users according to their needs.

After the GMET is constructed, we denote $G[VS]$ to be the graph induced by all vertices in the GMET from a single cleavage site of the peptide whose sequence we are trying to infer (termed a *basic structure*). From each $G[VS]$ an amino acid can be inferred. From connected basic structures determined from the GMET, continuous amino acids can be inferred. The detailed steps of inferring $G[VS]$ (and amino acids) from a GMET can be seen in [73]. Another graph, GMET_{AA}, is then used to represent the relationship of amino acid strings inferred from basic structures, in which the vertices V_{AA} and edges E_{AA} are defined as below [73].

V_{AA}: Each vertex represents an amino acid acquired from a basic structure of a GMET.

E_{AA}: Two amino acids are connected by an undirect edge if they are inferred from two connecting basic structures.

From a GMET, peptide sequences can be inferred by solving a path-finding problem in the simple graph GMET_{AA}. All paths are sorted in a decreasing order according to their length.

Since the GMET model considers multiple types of ions and different charges, the computational time and the possibility of false positives are greater than those with less sophisticated methods. In order to address this potential problem, more information from the spectra is integrated into the model. Peptide tags and AAC information are incorporated in this method as they were in previous work [73], but with adjustment for the ExD spectra. AAC, which consists of order-independent amino acid composition information of a peptide, is helpful [95, 96] in limiting the edges in a spectrum graph.

In the previous method [73], a single tag with a length of three was used to separate the peptide sequence into prefix, tag and suffix. Also an AAC in-house database was built containing all theoretically possible AACs for any given mass of no more than 3000Da [73]. However, since there are many ExD spectra produced from long peptides (typically ≥ 20), a single tag is not sufficient to simplify the sequencing problem, and the number of possible AACs increases dramatically with the increase of peptide length. Therefore, multiple length-3 peptide tags are considered here to separate the whole peptide into suitably sized parts for sequencing based on the GMET. The approach also yields a solution to the problem of increased numbers of possible AACs.

The results in [73] have shown that the single tag strategy is no longer effective when the peptide length is over 15. Based on the frequency of each amino acid occurring in a peptide [103], the average mass of a peptide of length 15 can be determined to be approximately 1600Da. Therefore, the threshold (as shown in Figure 4.1) to determine whether more tags are needed to separate a peptide is set to be $Thres = 1600$. With this value of $Thres$, the AAC database can be limited to masses of no more than 1600Da, which reduces the number of possible AACs considered in the GMET case dramatically.

An effective method named DirecTag [22] and the ranking criterion proposed in [73] are used here to generate and rank tags. In the following, detailed steps of integrating peptide tags and AACs are shown.

1. Generate length-3 tags using DirecTag [22] and put them into a set T according to their ranking scores in a decreasing order.

2. Select the first tag t, and delete t from T.

3. Calculate the mass of its prefix and suffix separated by this tag, denoted as Δm_{pre} and Δm_{suf}. Store the two mass intervals related to Δm_{pre} and Δm_{suf} into $Part$.

4. Go to Step 6 if Δm_{pre} and $\Delta m_{suf} \leq Thres$. Otherwise, find the mass interval(s) larger than $Thres$, denoted as $[m_s, m_e]$; go to Step 5.

5. Search in T for the tag whose two end ions are in the interval of $[m_s, m_e]$. If there is such a tag, denoted as $t^{'}$, go back to Step 3 using $t^{'}$ to separate $[m_s, m_e]$; else, go to Step 6.

6. Find out all possible AACs of mass intervals in $Part$ using the in-house AAC database.

7. Construct GMETs from both sides of the length-3 tag t. AAC information is applied to restrict the choices of edges in such GMETs. Specifically, only amino acids included in the AAC are considered when forming edges. With each possible AAC, one GMET is constructed.

8. Combine all partial sequences to be whole candidate peptides, and put them into set $Candidate$.

9. Go back to Step 2 if $T \neq \emptyset$, and use other tags in T; else, rank sequences in $Candidate$ and output results.

4.2.4 Candidate peptide ranking scheme

After all candidate peptides are put in set *Candidate* in the proposed algorithm, peptide ranking is the last step. Apart from the mass difference between parent peptide mass m_p and the candidate peptide mass (denoted as Δm_p), unique features in ExD spectra are considered. According to the literature [33, 87], x-ions (y-ion+CO) and z-ions (y-ion$-NH_3$, but not the $z^{\cdot}$-ion) are absent from ExD spectra. Therefore, if a candidate peptide P_c is the correct peptide that produced a specific spectrum S, x-ions and z-ions calculated from P_c should not be observed in S. That is to say, the fewer the number of x-ions and z-ions that are calculated from a candidate peptide sequence based on S, the more likely it is the correct peptide.

Therefore, in the proposed ranking scheme each candidate peptide P_c can be represented as a vector $CP(\Delta m_p, CZ_{match})$, where CZ_{match} is the total number of x-ions and z-ions calculated from P_c observed in S. The ranking scheme first orders CP according to Δm_p in a decreasing order, and then secondarily orders CP according to CZ_{match} in a decreasing order. After this process, the candidate peptides are output with their final ranking.

4.3 Experiments and Results

4.3.1 Datasets

Three ExD spectral datasets were used to investigate the performance of NovoExD. Another newly developed *de novo* peptide sequencing algorithm, pNovo+ [75], was used for comparison. It has been shown that pNovo+ achieves superior sequencing results on various testing data compared to another successful *de novo* sequencing software, PEAKS [62]. pNovo+ [75] can be used for HCD and (or) ExD spectra either alone or together. Here, the ExD option in pNovo+ is used.

The first dataset, SwedECD, contains ECD MS/MS spectra of doubly charged tryptic peptides [32]. The average length of peptides in the database is 10.6 residues, and the average mass is 1196.6Da. The other two datasets, SCX_ETDFT_no_decon and SCX_ETD_decon, are from the same research paper [98]. The original datasets contain various fragmentation MS/MS spectra including CID, HCD, and ETD spectra. The ETD spectra were selected for the experiments here. The SCX_ETD_decon dataset contains ETD spectra with deconvolution, and SCX_ETDFT_no_decon dataset contains raw data without deconvolution of spectra. For all spectra in these three datasets, a sequence is associated with each spectrum which is viewed as the correct sequence of the peptide producing the spectrum. The number and charges of spectra in each dataset are summarized in Table 6.2.

4.3.2 Parameters

There are several parameters needed for NovoExD, and the values applied in the experiments are listed in Table 6.3. θ, δ_1, δ_2, and the number of output sequence are set according to our previous study and

Table 4.3: Number of spectra and charges in each dataset used in the experiments

Dataset	Number of spectra	Charge of spectra
SwedECD	1414	+2
SCX_ETDFT_no_decon	1298	+2, +3, +4, +5
SCX_ETD_decon	612	+2, +3, +4, +5, +6

Table 4.4: Parameters used in the experiments

Parameter	Role in NovoExD	Value
Threshold θ	GMET edge building	0.01Da
δ_1	Charge determination	0.01Da
δ_2	Charge determination	0.01Da
$Thres$	Tag integration into GMET	1600Da
Number of tags	Tag integration into GMET	10 per spectrum
Number of output sequences	Candidate output	3 per spectrum

experiments [73, 100], and they can be changed by users for their needs. The value of $Thres$ has been introduced previously in detail. The number of tags is chosen to be 10 because that the tags ranked lower than the top 10 tags are most likely to be wrong tags according to our previous study [73].

4.3.3 *De novo* peptide sequencing performance

All three datasets were used to investigate the performance of NovoExD and pNovo+. For each spectrum in a dataset, the top three candidates produced by the two methods were output. If any one of the three candidates from a spectrum was correct (the same as the sequence associated with this spectrum), it was deemed that full length accuracy was achieved for this spectrum and for the given method. In order to analyze the contribution of multi-tags and ion charges in NovoExD, sequencing results of using the GMET model without multi-tags and ion charge information are also shown. Detailed comparison is presented in Table 4.5. Since the SwedECD dataset consists of +2 spectra with average peptide length of 10.2 [32], most of the fragment ions have charge +1 and a single tag is sufficient for peptide sequencing. Therefore, we did not use multi-tags and ion charges in the comparison and recorded "N/A" in the associated box in Table 4.5; using these features is expected to yield results similar to those shown.

From Table 4.5 it is evident that the GMET model achieves better performance than pNovo+ on the SwedECD dataset. Results for two datasets consisting of highly charged ($\geq +3$) spectra, SCX_ETDFT_no_decon

Table 4.5: Average full length peptide sequencing accuracy comparison

Dataset	pNovo+	GMET without multi-tags and ion charge information	Proposed NovoExD
SwedECD	52.86%	**72.25%**	N/A
SCX_ETDFT_no_decon	63.02%	64.10%	**85.83%**
SCX_ETD _decon	64.62%	59.15%	**88.70%**

and SCX_ETD_decon, allow us to see the potential contribution of multi-tags and ion charges to peptide sequencing. The results in Table 4.5 illustrate that NovoExD obtains the highest full length accuracy among all three methods, and the GMET without multi-tags and ion charges has performance similar to that of pNovo+. Therefore, using these features dramatically improves the identification accuracy of highly charged spectra ($\geq +3$). Specifically, accuracy is improved 22.81% and 24.08% compared to pNovo+, and 21.73% and 32.55% compared to the GMET model without multi-tags and ion charges.

Furthermore, we considered the relationship between the number of correctly identified peptides and peptide length. Figures 4.2 and 4.3 summarize the results from comparing the output of NovoExD and pNovo+ on two datasets, SCX_ETDFT_no_decon and SCX_ETD_decon, respectively. These figures show that NovoExD outperforms pNovo+ on almost every peptide length except lengths 15 and 19 on SCX_ETDFT_no_decon, and lengths 12 and 14 on the SCX_ETD_decon dataset, respectively. For length greater than 16, NovoExD achieves almost perfect identification on the SCX_ETD_decon dataset, while pNovo+ has lower identification accuracy. The lower accuracy of NovoExD for some lengths may be due to the threshold controlling whether multi-tags are used. Some peptides may require multi-tags for better performance but do not achieve the pre-defined threshold. In future study, we will explore more suitable thresholds for the use of multi-tags.

In addition, the relationship between the number of correctly identified peptides and spectrum charge was examined. Tables 4.6 and 4.7 summarize the results of the comparison between NovoExD and pNovo+ on the SCX_ETDFT_no_decon and SCX_ETD_decon datasets, respectively. These tables show that NovoExD outperforms pNovo+ under every spectrum charge. Thus for the two datasets consisting primarily of +3 and +4 charged spectra, NovoExD identifies more peptides, which argues for the potential contribution of using ion charge information for MS/MS peptide sequencing, especially with higher charged spectra. Since the experiment here is limited, more investigation is needed to make a conclusive statement.

Finally, we examine the computational time of NovoExD. The algorithm was written using MATLAB (2010b) and the program was run on a PC with a 3.07 GHz quad-core CPU and MS Windows 7 operating system. The results for different datasets are shown in Table 4.8. Since a charge determination step is part of NovoExD, in order to examine the effect of this step on the total running time, the time without charge determination is also calculated for the SCX_ETDFT_no_decon and SCX_ETD_decon datasets. The

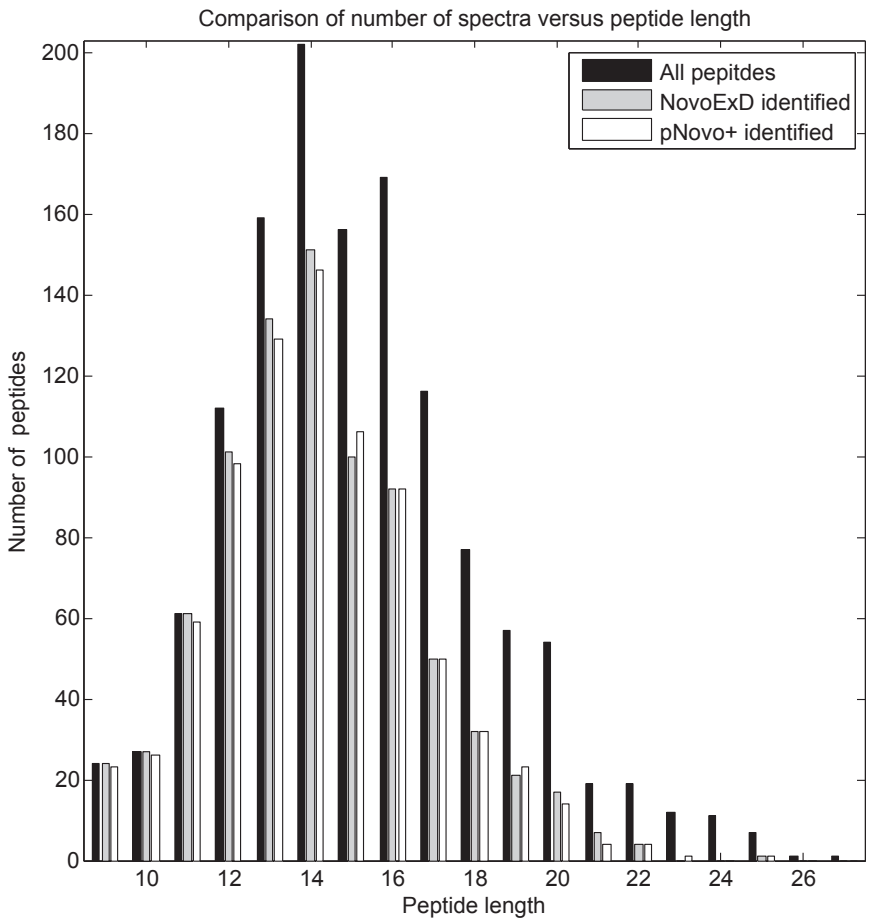


Figure 4.2: Comparison of the number of correctly identified peptides and peptide length between NovoExD and pNovo+ using the SCX_ETDFT_no_decon dataset.

SwedECD dataset contains only +2 charged spectra and all ions are assumed to be in charge state +1. Therefore, there is no charge determination step needed for SwedECD dataset, and "N/A" is recorded in the associated table cell. All run times (CPU times) in this table are given in seconds per spectrum.

From Table 4.8 one can see that the computational time varies among different datasets, but are all in an acceptable time range; specifically, the running time per spectrum is less than 2 seconds. Another interesting result is that the running time without charge determination is far less than that with charge determination. This step takes around two thirds of the total time. Therefore, in order to enhance the speed of NovoExD, the focus should be on this most time-consuming step. If a faster and more efficient charge determination method is developed, the whole running time can be reduced dramatically. Such investigation will be part of our future study of this *de novo* peptide sequencing approach.

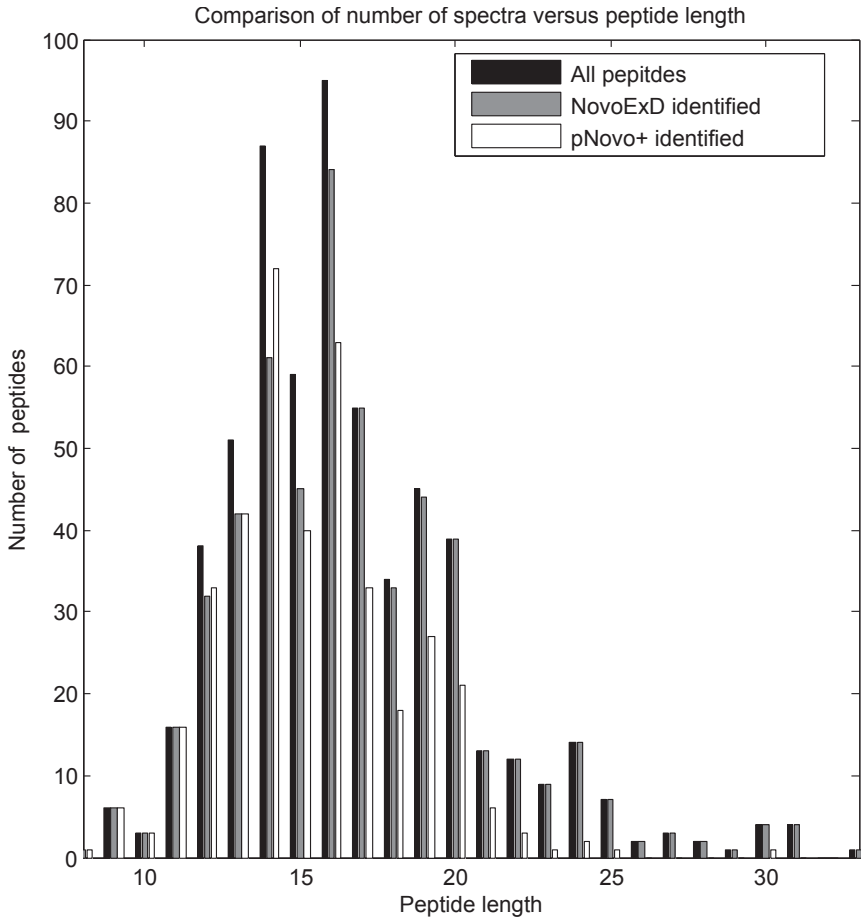


Figure 4.3: Comparison of the number of correctly identified peptides and peptide length between NovoExD and pNovo+ using the SCX_ETD_decon dataset.

Table 4.6: Comparison between the number of correctly identified peptides and spectrum charge using the SCX_ETDFT_no_decon dataset

Spectra charge	pNovo+ identified	NovoExD identified	Total spectra
+3	488	583	682
+4	311	457	532
+5	1	9	9

Table 4.7: Comparison between the number of correctly identified peptides and spectrum charge using the SCX_ETD_decon dataset

Spectra charge	pNovo+ identified	NovoExD identified	Total spectra
+3	181	204	247
+4	199	295	320
+5	5	28	28
+6	0	2	2

Table 4.8: Running time comparison on different datasets using NovoExD

Dataset	SwedECD	SCX_ETDFT_no_decon	SCX_ETD_decon
Number of Spectra	1414	1298	612
Time (sec.) per spectrum for NovoExD	0.41	2.09	0.53
Time (sec.) per spectrum without charge determination	N/A	0.45	0.18

4.4 Conclusions and future work

In this paper, a *de novo* peptide sequencing method for ExD spectra, NovoExD, has been proposed. It uses a new spectrum graph model, considers multiple peptide tags to separate a peptide into small mass regions, and integrates fragment ion charge and amino acid composition (AAC) information. Then, it combines small regions to output complete sequences of candidate peptides. In addition, a charge determination step is used to extract more information from highly charged ExD spectra.

Three datasets of ExD spectra were used to investigate the performance of NovoExD by comparing it with another successful *de novo* peptide sequencing method, pNovo+, which has an option for ExD spectra. Experimental results have shown that the improvements in terms of average full length peptide sequencing accuracy are over 20% on all datasets when compared to pNovo+.

In future, we will improve NovoExD's accuracy, reduce its computational time, expand the evaluation of NovoExD to more ExD datasets, and optimize the parameters in NovoExD. In addition, since a popular approach to peptide sequencing is pairing different types of spectra from the same peptide to obtain more information, we are planning to modify the model and apply it to the multiple spectra sequencing problem.

Acknowledgment

This work was supported by Natural Sciences and Engineering Research council of Canada (NSERC).

Addendum

In the Subsection 4.2.2, when determining the ion charges, the equation in step 3 is from the previously introduced equations 4.2 and 4.3.

Errata

1. In the second paragraph of Section 4.2, "The motivation of ..." should be changed to "The motivation for ..."
2. In the forth paragraph of Section 4.2, "else," should be changed to"otherwise,".
3. In the fifth paragraph of Section 4.2, "These steps in the algorithm and the various variables ..." should be changed to"These steps in the algorithm and the different variables ...,".
4. In the Subsection 4.2.2, when determining the ion charges, the requirement of "where $\xi_p + \xi_q = n$" in step 3 should be removed.

Chapter 5

A framework of *de novo* peptide sequencing for multiple tandem mass spectra

Published as: Yan Yan, Anthony J. Kusalik and Fang-Xiang Wu. "A framework of *de novo* peptide sequencing for multiple tandem mass spectra," IEEE Transactions on NanoBioscience, vol.14, no.4, pp.478-484, June 2015.

In the previous chapter, a new *de novo* peptide sequencing method for ETD/ECD spectra, NovoExD, is presented. NovoExD modifies the model in NovoHCD and adds a charge determination step for multiple charged ECD/ETD spectra. With the availability of various MS/MS spectra, the use of multiple spectra from the same peptide for sequencing has become a new, popular research topic.

Multiple spectra peptide sequencing has the potential to extract more information from spectra than single-spectrum based sequencing since information missed from one spectrum may be found in others. This advantage thus gives multiple spectra sequencing the potential to significantly increase the accuracy and practicality of *de novo* sequencing. Currently, there are some methods designed for multiple spectra sequencing, but the models may not be designed well enough to extract as much information as they could. There is still room to develop advanced methods in this area. Therefore, the study presented in this chapter focuses on multiple spectra sequencing.

This chapter presents a framework for multiple spectra sequencing and applies it to paired CID (or HCD) and ECD (or ETD) spectra. The newly available MS/MS has two fragmentation modes installed, for example, the CID and ETD modes. The two models are changed very quickly during the experiment so that the two types of spectra output at the same time are viewed as the ones generated by the same peptide. These spectra have different dominant fragment ions and are complementary to each other. They are the most widely used pairs in MS/MS based peptide sequencing. The performance of the framework with application to this kind of spectra pair is compared to another competing method named pNovo+. Experiential results on several dataset pairs show that the proposed framework outperforms pNovo+ in terms of full length sequencing accuracy on experimental datasets.

Abstract

With tandem mass spectrometry (MS/MS), spectra can be generated by various fragmentation techniques including collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), electron capture dissociation (ECD), electron transfer dissociation (ETD) and so on. At the same time, *de novo* sequencing using multiple spectra from the same peptide generated by different fragmentation techniques is becoming popular in proteomics studies. The focus of this study is the use of paired spectra from CID (or HCD) and ECD (or ETD) fragmentation because of the complementarity between them. We present a *de novo* peptide sequencing framework for multiple tandem mass spectra, and apply it to paired spectra sequencing problem. The performance of the framework on paired spectra is compared to another successful method named pNovo+. The results show that our proposed method outperforms pNovo+ in terms of full length peptide sequencing accuracy on three pairs of experimental datasets, with the accuracy increasing up to 13.6% compared to pNovo+.

5.1 Introduction

When dealing with mass spectrometry-based peptide sequencing problems, there are three main kinds of methods used: database searching, peptide tagging and *de novo* sequencing [17]. In database searching, theoretical spectra are computed from an existing protein database and peptides are identified by matching the theoretical spectra to experimental spectra [46]. The major disadvantage of database searching is that it cannot identify new or unknown peptides. Peptide tagging [21, 22] is usually used to reduce the search space and time, and has the potential to improve *de novo* peptide sequencing. *De novo* sequencing has the ability to identify new proteins, proteins resulting from mutations, proteins with unexpected modifications and so on. With the recent developments of high mass-accuracy MS/MS and alternative fragmentation techniques, *de novo* sequencing has shown promising developments [48]. Therefore, this study focuses on MS/MS *de novo* peptide sequencing.

Tandem mass spectrometry (MS/MS) is a commonly used technology for peptide sequencing. It measures the mass-to-charge ratio (m/z) of the components in experimental compounds, and the data collected from it is called a tandem mass spectrum (MS/MS spectrum) [48]. In MS/MS, peptide ions are fragmented into different kinds of fragment ions, named a-, b-, c-, x-, y-, and z-ions. Different fragmentation techniques used in MS/MS yield differing types of dominating fragment ions. Collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD) yield b-ions and y-ions as dominating ions [36]. Electron capture dissociation (ECD) and electron transfer dissociation (ETD) preferentially produce variants of c-ions and z-ions, and occasionally a-ions [31–33]. ETD [37] is a modification of the ECD technique [38]; it produces high quality MS/MS spectra for multi-charged peptides, has no strong cleavage preferences, and preserves labile post-translational modifications (PTMs) [41]. These features yield spectra containing more

useful information, which has the potential to give better performance when used in peptide sequencing. In addition, all the fragment ions usually lose some small molecules such as H_2O and NH_3 during the fragmentation.

With the appearance of alternative MS/MS spectra resulting from different fragmentation techniques, novel computational methods have emerged to enhance *de novo* peptide sequencing performance. For example, pNovo [72], which applies a spectrum graph model and combines immonium ions (IMs) and internal fragment ion information from HCD spectra, has achieved superior peptide sequencing results. Its performance on various testing data has been shown to be better than that of two previous algorithms, PEAKS [62] and PepNovo [59]. We have previously proposed a *de novo* sequencing method for ECD and ETD spectra [17] based on a new type of graph model, and the experiments on several MS/MS datasets showed that this method achieved better peptide sequencing results, compared to another similar method [17] .

Apart from the algorithms based on single spectrum peptide sequencing, some researchers have tried to use multiple spectra from the same peptide to infer peptide sequences [31, 75, 77, 79, 81]. Multiple spectra peptide sequencing is promising because it has the potential to extract more information from spectra, and significantly increase the accuracy and practicality of *de novo* sequencing. The use of paired CID (or HCD) and ECD (or ETD) spectra is the major focus of the current study because of the availability and complementary properties of these spectra.

Savitski *et al.* [79] first proposed a computational method that used two spectra from the same peptide to infer peptide sequences. In their method, peaks appearing in both spectra representing the same partial peptides were considered to be much more reliable than the rest of the peaks; they were used to create a backbone of the sequence. The less reliable peaks were then used to fill the gaps in the sequence backbone or extend the sequence until a full sequence was obtained. Another algorithm, CompNovo [77], employed a divide-and-conquer approach combined with a mass decomposition algorithm, and extracted information from CID and ETD fragmentation for peptide sequencing. This algorithm showed better peptide sequencing results compared to the ones that only used CID spectra. In 2010, He and Ma [31] presented a new algorithm, ADEPTS, to utilize multiple spectra for *de novo* sequencing. It was mainly focused on a new scoring function. A new way of using intensity information was also included in this algorithm, which achieved better results than programs like CompNovo. Most recently, Chi *et al.* proposed a *de novo* sequencing method named pNovo+ [75] for HCD and ETD spectra pairs. It applied a spectrum graph model and combined different fragment ion information from the two spectra, and showed superior results on various testing data.

In this paper, we present a new framework to deal with the multiple spectra sequencing problem, and test its performance using spectra pairs from the same peptides. The proposed framework builds a modified spectrum graph with multiple edge types (GMET) [73] by considering frequently observed fragment ions from all spectra, and utilizes unique features from these spectra. Amino acid compositions and peptide tags from both spectra are incorporated into the proposed method with consideration of the multiple spectra situation.

5.2 Methods

In this section, the framework is proposed. It first considers unique features from all experimental spectra to generate a merged spectrum S_m, and then uses peptide tags to break a whole peptide sequence into smaller regions. After that, the proposed method builds a GMET model for each region and uses amino acid composition (AAC) information to limit the edge numbers in the GMET. It finally combines all smaller parts to output complete sequences of candidate peptides. The whole framework is summarized in Figure 5.1.

5.2.1 Basic ion types considered in each spectrum

Previous study has determined the frequency of different fragment ions observed in different MS/MS spectra [17, 32, 33, 41, 73]. The ions listed in Tables 6.1 and 5.2 are considered in the proposed method based on the availability of spectra and observed frequency of different ions. At this time, CID, HCD, ECD, and ETD are the most popular spectra used. If more kinds of spectra are available in future, new types of ions could be incorporated into the framework.

In these tables, m_{H_2O}, m_{NH_3}, m_H, m_{CO} denote the mass of H_2O, NH_3, H, and CO- groups, respectively. $\sum(residue\ mass)$ is the mass sum of all amino acids from an end amino acid of a peptide sequence to the amino acid at the current cleavage site. Then, $\sum(previous\ residue\ mass)$ equals $\sum(residue\ mass)$ taking off the mass of the amino acid at the current cleavage site. b_m and y_m are the masses of the b-ion and y-ion at the current cleavage site, respectively. $x_{previous}$ is the mass of the x-ion previous to the current one. An ion annotated with "$\bullet$" is a radical fragment ion.

In Tables 6.1 and 5.2, the first column presents the types of ions considered, and the second and third columns give two kinds of mass calculation. The second column is the calculation based on previous residues, which gives the relationship among cleavage sites. The third column is the calculation based on other types of ions (specifically, the b- and y- ions) at the same cleavage site.

Since ions lose small molecules and some fragmentation techniques yield special ions during MS/MS experiments, the proposed method considers the following ions: basic types of ions in Tables 6.1 and 5.2, small molecule loss (such as H_2O and NH_3) from all the basic types of ions, multi-charged (charges up to $n-1$ for charge n precursor ion) basic ion types, and immonium ions (IMs) occurring frequently in HCD spectra.

5.2.2 Spectra merging

The basic idea of spectra merging is to select signal ions from each spectrum to form a new spectrum, denoted as S_m, which contains more useful information. Here, two kinds of relationships between ions are considered: amino acid mass difference and complementarity [90].

For amino acid difference, with consideration of regular amino acid masses and loss of H_2O and NH_3, all 2-tags (two amino acids long) in each spectrum are first produced. Here, a length-2 tag consists of three

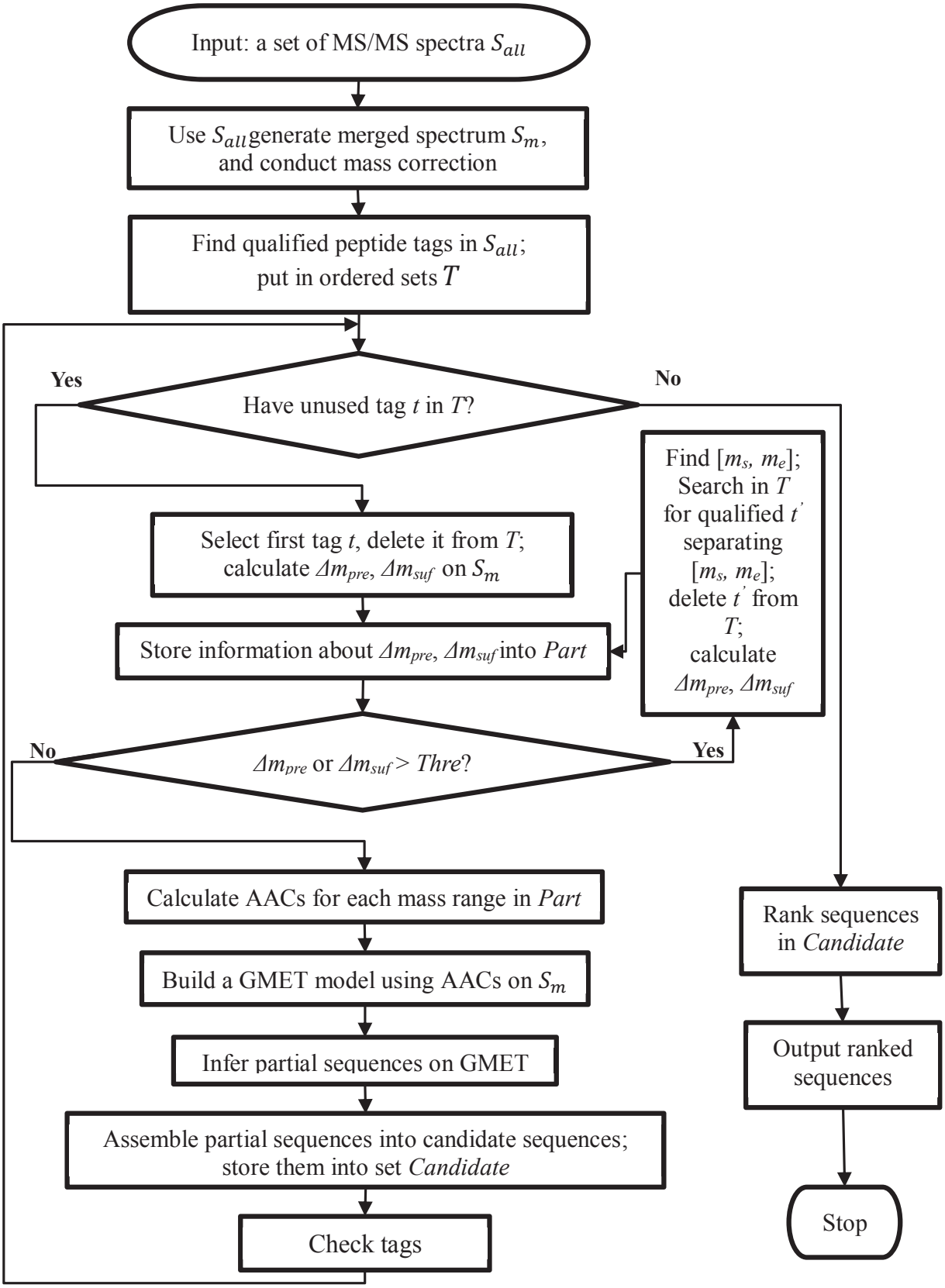


Figure 5.1: A flow chart of the proposed framework.

Table 5.1: Ion types considered in CID/HCD spectra

Ion Type	Mass calculation from residues	Mass calculation from other ions
a	$\sum(residue\ mass) - 26.9871$	$b_m - m_{CO}$
b	$\sum(residue\ mass) + 1.0078$	b_m
x	$\sum(residue\ mass) + 44.9977$	$y_m + m_{CO}$
y	$\sum(residue\ mass) + 19.0814$	y_m

Table 5.2: Ion types considered in ECD/ETD spectra

Ion Type	Mass calculation from residues	Mass calculation from other ions
c	$\sum(residue\ mass) + 18.0344$	$b_m + m_{NH_3}$
$c-1^{\cdot}$	$\sum(residue\ mass) + 17.0265$	$b_m + m_{NH_3} - m_H$
$z^{\cdot}$	$\sum(residue\ mass) + 3.0156$	$y_m - m_{NH_3} + m_H$
$z+1$	$\sum(residue\ mass) + 4.0156$	$y_m - m_{NH_3} + 2m_H$
w	$\sum(previous\ residue\ mass) + 73.0290$	$x_{previous} + m_{CO}$
b	$\sum(residue\ mass) + 1.0078$	b_m
y	$\sum(residue\ mass) + 19.0814$	y_m

ions (any two consecutive ion pairs having mass difference close to an amino acid mass or an amino acid mass plus/minues some small molecular) from an experimental spectrum. The three consecutive ions can be denoted as u, v, t in spectrum S. Without loss of generality, we say the masses of u, v, t are in an increasing order. Then, ion v is named as a "middle ion" and selected for the merged spectrum. The middle ion, which has two other supporting ions from the spectrum, is more likely to be a signal ion rather than noise. For the complementary relationship, all complementary ion pairs in each spectrum are selected.

For an experimental spectrum S and ions $u, v, t \in S$, the charges of the ions are denoted ξ_u, ξ_v, and ξ_t, and the m/z values of these ions in charge state $+1$ are denoted as u^+, v^+, and t^+. Since charges of ions are unknown, in order to calculate all possible m/z values, we assume ξ_u, ξ_v, and ξ_t to be any positive integers from $+1$ to $(n-1)$, and calculate the associated $n-1$ spectra having all ions with charge state $+1$, denoted as $S_{\xi_i to 1}$. Here n is the charge value of a given spectrum. There are already many algorithms for MS/MS spectrum charge determination [92,93], and here we use a straightforward one similar to that in [17] to simplify the calculation.

Additionally, we denote A as the set of 20 amino acids, and $a_i \in A$ as a certain amino acid. a_i is also used to represent its residue mass. m_{loss} is defined to be the mass of some small molecules or groups lost

Table 5.3: Relationships and ions selected in spectra merging.

Relationship	Ions selected
$\mid (v^+ - u^+) - a_i + \sigma \mid \leq \theta$ and $\mid (t^+ - v^+) - a_j - \sigma \mid \leq \theta$	v (middle ion)
$\mid m_p + 2m_H - (v^+ + u^+) \pm \sigma \mid \leq \theta$	v and u if $u, v \in S_c$
$\mid m_p + 3m_H - (v^+ + u^+) \pm \sigma \mid \leq \theta$	v and u if $u, v \in S_e$

from fragment ions, which includes H_2O and NH_3. Relationships and ions in Table 5.3 are then utilized. In the table, $a_i, a_j \in A$; σ can be 0 or m_{loss} (considering the loss of small molecules of fragment ions); θ is a given threshold; and m_p is the parent peptide mass. Finally, S_m consists of all middle ions in 2-tags and complementary ion pairs for a given pair of MS/MS spectra. All ions in S_m come with charge values from this step.

Here, we give a simple example to show how the merging step works. Assume the m/z values of two experimental spectra are $S_c = \{130, 199, 277, 346\}$ (represent a CID spectrum) and $S_e = \{132, 182, 234\}$ (represent a ETD spectrum), respectively. The parent mass is $m_p = 492$. S_c is in charge state +2 and S_e is in charge state +3. The lost small molecule is H_2O, and then $m_{loss} = 18$. Here, we use integers to simplify the calculation and focus on the principles of the method.

Converting all ions to charge state +1, three associated spectra are generated:

$$S^c_{1to1} = \{130, \underline{\mathbf{199}}, \underline{277}, 346\},$$
$$S^e_{1to1} = \{\underline{132}, \underline{182}, 234\},$$
$$S^e_{2to1} = \{263, 363, 467\}.$$

We first deal with S^c_{1to1}. From the calculations in Table 5.3, we get that values 130, 199, and 346 satisfy $\mid (199 - 130) - a_S + 18 \mid = 0$ and $\mid (346 - 199) - a_E - 18 \mid = 0$, where $a_S = 87$ and $a_E = 129$ are the masses of serine and glutamine, respectively. Then we infer that the ion having m/z value of 199 (in boldface above) is a fragment ion having lost a molecular of water, and it is added to the merged spectrum S_m. In addition, we get that values 199 and 277 (underlined above) satisfy $\mid 492 + 2m_H - (199 + 277) - 18 \mid = 0$. Then we infer that these two ions are complimentary ions, and the ion having m/z value of 277 is a regular ion. Both ions are in charge state +1.

We now deal with S^e_{1to1} and S^e_{2to1}. Values 132 and 363 (underlined above) satisfy $\mid 492 + 3m_H - (132 + 363) \mid = 0$. Then we infer that these two ions are complimentary ions, and the ion having m/z value of 182 is in charge state +2 (the ion at the same position as ion 363 in S^e_{1to1}). Therefore, the final $S_m = \{132, 182, 199, 277\}$, and their charges are known from the above steps.

5.2.3 Parent mass correction

In this problem there are multiple MS/MS spectra and each comes with a parent peptide mass. These masses are probably different, though they should be close. Therefore, before *de novo* sequencing, parent mass correction is needed. Here, complementary ion pairs are used to find the optimal parent mass in a given region. The model hypothesis is that the real parent mass has the minimal mass difference to complementary ion pair masses.

Denote complementary ion pairs in S_m as $CIP = \{(I_j, I_j^c) \mid j = 1, 2, \cdots k\}$, where $I_j, I_j^c \in S_m$ are complementary ions, and k is the total number of complementary ion pairs. We find the optimal parent mass P_{mass} by solving the following optimization problem

$$\begin{aligned} &\min \Sigma_{j=1}^{k} \mid (I_j + I_j^c) - P_{mass} \mid^2 \\ &s.t. \;\; P_{inf} \leq P_{mass} \leq P^{sup} \end{aligned} \tag{5.1}$$

Here, P_{inf} and P^{sup} are the infimum and supremum of P_{mass}, respectively. In practice, users could set the two masses as the maximum and minimum masses of these spectra, or other suitable values as needed. Since the above problem is a convex optimization problem with one unknown variable P_{mass}, the minimal value of P_{mass} can be uniquely determined [104].

5.2.4 *De novo* sequencing model

The sequencing model proposed here is derived from [17] and [73] with modifications designed for the multiple spectra sequencing problem. It uses a new type of spectrum graph with multiple edge types (GMET), considers multiple peptide tags to separate a peptide into small mass regions, and integrates fragment ions in the merged spectrum S_m and amino acid composition (AAC) information. For each mass region combination (see Figure 5.1, set $Part$) separated by several tags, a GMET is built.

In graph $GMET = (V, E, \Xi)$ each peak (corresponding to a fragment ion) in the experimental spectrum is represented as a vertex $v \in V$ and its m/z value is denoted as $(m/z)_v$. Each v has a charge value $\xi \in \Xi$, which is determined by the previous merging step. The default value of ξ is $+1$. $\forall u, v \in V$, let m_u and m_v denote their mass values, then we have $m_u = ((m/z)_u \times \xi_u) - (\xi_u - 1)$ and $m_v = ((m/z)_v \times \xi_v) - (\xi_v - 1)$. Five different types of edge in E are considered in the GMET, and the detailed calculations can be found in [17].

A major difference in this GMET compared to the one in [17] is that it considers all ion types in Tables 6.1 and 5.2 when sequencing using S_m. Another difference in the sequencing model is the use of peptide tags. Since there are multiple experimental spectra and one merged spectrum S_m, instead of generating tags from S_m, the model uses experimental spectra separately to produce multiple peptide tag sets (by the DirecTag algorithm [22]), and uses the ranking scheme in [73] to sort them. The reason for this is that DirecTag is

designed for a single MS/MS spectrum, not a merged spectrum, and the mixed types of ions in S_m are a source of difficulty for DirecTag.

With more types of ions considered, the computational time and the possibility of false positives may be increased in the model. As a solution, AAC, which consists of order-independent amino acid composition information of a peptide [95], is incorporated into the model to limit the edges in a GMET. For a given mass region, possible AACs are generated for such specific mass; and only amino acids in an AAC will be used as edge choices in a GMET. From the GMET model, the proposed method infers partial peptide sequences on each segment, and assembles them into final candidate sequences. The steps of sequencing by the GMET model with integrated peptide tags and AAC are shown in Figure 5.1.

5.2.5 Candidate peptide ranking

When all candidate peptides have been produced, the last step is to rank these candidates and determine the most likely correct ones. A theoretical spectrum generated from the correct sequence is expected to have the best match to the experimental spectrum. Therefore, we use this feature as well as ion intensity to design a candidate ranking scheme. The detailed steps are shown below.

1. For a candidate peptide sequence $P_{can} \in Candidate$, generate a theoretical spectrum for each experimental spectrum. Ion types considered are the ones in this specific spectrum that were used for generating spectrum S_m. All ion intensities in the theoretical spectra are defined as a constant value.

2. Compare the theoretical spectra with experimental spectra, and select matching ions from each pair of spectra with a mass difference less than the predefined threshold θ_r.

3. Sum the intensity values of matching ions in the experimental spectra to be the final ranking score of P_{can}.

After this process, each candidate peptide will be assigned a ranking score. The higher the score, the more likely it is that this sequence is the correct one generating the input experimental spectra. In the proposed method, for a set of input spectra, the top 3 highest scoring candidates along with their ranking scores are output as results.

5.3 Experiments and Results

We applied the framework to MS/MS spectra pairs to evaluate its performance. Here, three pairs of datasets were used. Another newly developed *de novo* peptide sequencing algorithm for paired spectra, pNovo+ [75], was used for comparison. pNovo+ [75] can be used for HCD and (or) ETD/ECD spectra either alone or together. It has been shown that pNovo+ achieves superior sequencing results for paired spectra sequencing on various testing data compared to other methods [75]. In addition, our previously proposed methods for

Table 5.4: Number of spectra and charges in each dataset used in the experiments

Dataset	Number of total spectra	Charge of spectra	Number of selected spectra
SwedECD	11491	+2	3119
SwedHCD	10878		
SCX_HCD_decon	1952	+2 to +6	402
SCX_ETD_decon	612		
SCX_HCD_no_decon	2557	+2 to +5	753
SCX_ETD_no_decon	1298		

HCD and ECD/ETD spectra [17,73] were included for further comparison. The detailed experimental process and results comparison are presented below.

5.3.1 Datasets

There are three pairs of HCD and ECD/ETD spectral datasets used in the experiments. The first pair of datasets are the SwedHCD and SwedECD datasets, which contain doubly-charged MS/MS spectra of unique peptides [32, 97]. The other two pairs of datasets, SCX_HCD_decon and SCX_ETD_decon, plus SCX_HCD_no_decon and SCX_ETD_no_decon, are from the same research paper [98]. The latter dataset pair (labeled with "_no_decon") contains raw data without deconvolution of spectra while the other pair contains spectra with deconvolution. The original datasets contain various fragmentation MS/MS spectra including CID, HCD, and ETD spectra. The HCD and ETD spectra were selected for the experiments here. The reason that HCD instead of CID spectra were selected is that, typically, HCD spectra are more informative and contain special ions (immonium ions). Each spectrum in the above datasets has a correct sequence associated with it. In order to conduct our experiments, we selected spectra pairs having the same peptide sequence from the paired datasets. The number of spectra, the charges of spectra, and the number of selected pairs of spectra in the datasets are summarized in Table 6.2.

5.3.2 *De novo* peptide sequencing performance

Before investigating the performance of the proposed method, we first conducted peptide sequencing based on a single spectrum. There was no need to test the proposed method on the spectra pairs for which single spectrum methods output identical and correct results. Such spectra are of sufficient quality that any one of the two produced a satisfying result. The remaining paired spectra constitute a much more rigorous test and those were used in testing the proposed method and comparing it against the other sequencing algorithms.

Table 5.5: Full length peptide sequencing accuracy based on a single spectrum for different datasets.

Dataset	Method	Accuracy	Number of correctly identified spectra by both methods
SwedHCD	NovoHCD	95.80% (2988)	2516
SwedECD	NovoGMET	86.47% (2697)	
SCX_HCD_decon	NovoHCD	80.59% (324)	241
SCX_ETD_decon	NovoGMET	87.31% (351)	
SCX_HCD_no_decon	NovoHCD	84.99% (640)	504
SCX_ETD_no_decon	NovoGMET	85.66% (645)	

Since our previously proposed single spectrum sequencing methods NovoHCD [73] and NovoGMET [17] both showed superior results on various datasets compared to pNovo and pNovo+, sequencing based on these two methods was conducted. For each spectrum, the top three candidates output were considered. If any one of the three candidates interpreted from a spectrum was correct, we say that the method achieved a full length accuracy for the given spectrum. The results for full length accuracy based on a single spectrum are presented in Table 5.5. In this table, the numbers in brackets are the counts of correctly identified spectra. The last column of this table is the number of correctly identified spectra by both methods for a given dataset pair. For instance, in datasets SwedHCD and SwedECD, single spectrum sequencing methods NovoHCD and NovoGMET identified 2988 and 2697 spectra, respectively. Among those spectra, 2516 spectra were correctly identified by both methods, which means that NovoHCD and NovoGMET output identical and correct sequencing results for these spectra.

Then, we filtered out those spectra pairs that both NovoHCD and NovoGMET have correctly identified, and used the remaining spectrum pairs to compare the performance of the proposed method and pNovo+. The total number of spectrum pairs selected from the three dataset pairs were 3119, 402 and 753, respectively. The numbers of spectrum pairs correctly identified by the two single spectrum sequencing methods were 2516, 241 and 504, respectively. Then the remaining spectrum pairs for the proposed method versus pNovo+ performance comparison were 603, 161 and 249, respectively. These numbers are shown in Table 5.6 together with the full length peptide sequencing results for the proposed method and pNovo+.

From Table 5.6 one can see that for all datasets, the proposed method achieved higher full length accuracy than pNovo+, and the improvement is up to 13.6%. The improvement varies indicating property differences among these datasets. The first dataset pair produced the highest accuracy, and this may be partly because

Table 5.6: Full length peptide sequencing accuracy comparison among different datasets.

Dataset	Number of spectra pairs	accuracy of pNovo+	accuracy of the proposed method
SwedHCD and SwedECD	603 (=3119-2516)	81.76% (493)	**95.36%** (575)
SCX_HCD_decon and SCX_ETD_decon	161 (=402-241)	77.02% (124)	**83.85%** (135)
SCX_HCD_no_decon and SCX_ETD_no_decon	249 (=753-504)	84.74% (211)	**94.78%** (236)

of the low spectrum charge (+2) in this pair. Low spectrum charge results in the fragment ions in such spectra having lower charges and less complexity than the ions in higher charged spectra. This makes the spectra simpler and easier for the methods to process and interpret correctly. Similarly, we see that the second dataset pair contains the largest spectrum charge range (from +2 to +6), and both pNovo+ and the proposed method achieved their lowest sequencing accuracy on it.

Furthermore, we considered the relationship between the number of correctly identified peptides and peptide length. Comparison between the proposed method and pNovo+ on SwedHCD and SwedECD datasets is shown in Figure 5.2. The reason for showing this pair is because it has the largest number of spectra among all datasets, and is expected to provide the most comprehensive comparison.

Figure 5.2 shows that the proposed method outperforms pNovo+ on every peptide length in this pair of datasets. The sequencing results of the two methods are similar when peptide length is small – for example, less than 9 – and both methods achieve almost perfect results. However, with increased of peptide length, the proposed method shows significantly better results than pNovo+. One can see that when the peptide length is 21, pNovo+ identifies no correct results while the proposed method identified all spectra in this pair of datasets. Another interesting result from the Figure 5.2 is that when the peptide length is 22, both methods identify all spectra. The varied performance of pNovo+ may indicate the complexity of the spectra generated by long peptides. However, since the number of such spectra in this pair of datasets is small, further experiments and analysis of spectral properties are needed to get a firm conclusion.

The remaining two dataset pairs in the experiments contain multiple charged spectra. Therefore, they can be used to examine the relationship between the number of correctly identified peptides and peptide charge. Since dataset pair SCX_HCD_no_decon and SCX_ETD_no_decon contains more spectrum pairs, this dataset pair is used. Figure 5.3 summarizes the comparison between the proposed method and pNovo+. Since two spectra in a pair may have different charge values, the higher charge is used as the spectrum charge of this

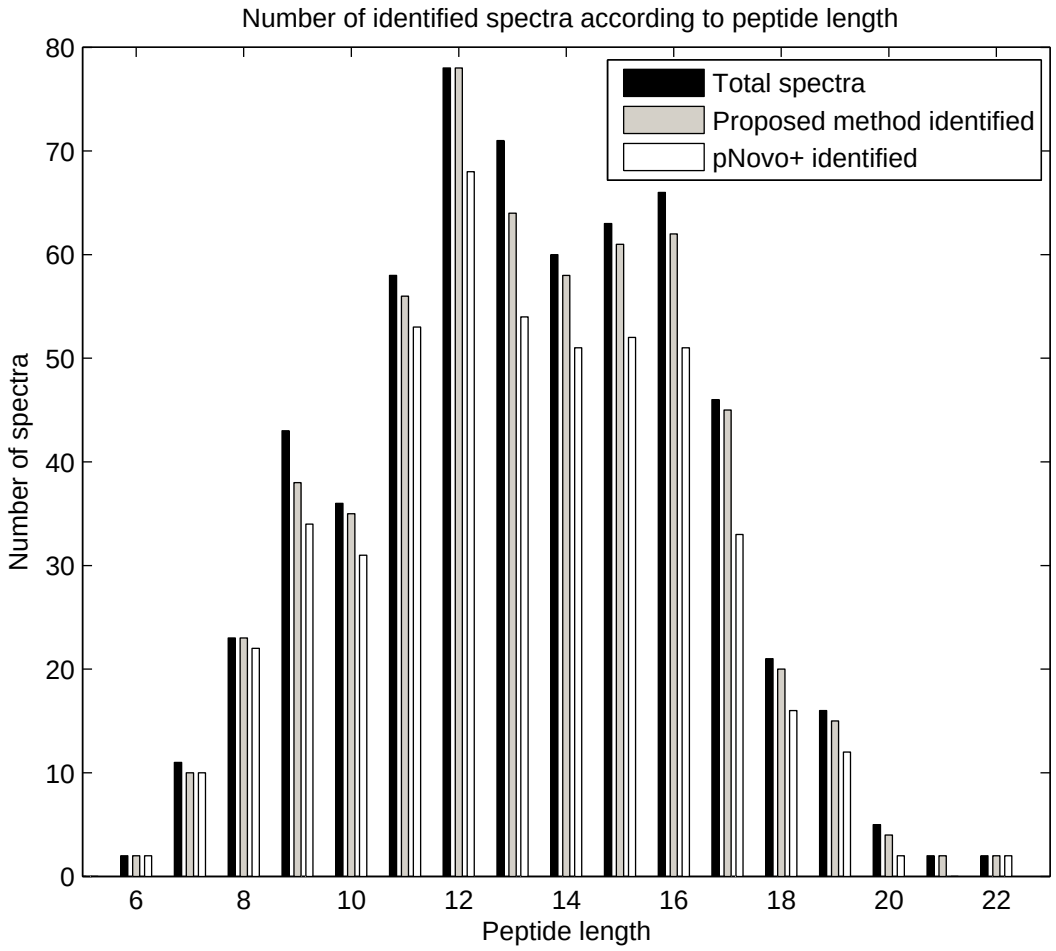


Figure 5.2: Comparison of the number of correctly identified peptides verses peptide length for the proposed method and pNovo+ on the SwedHCD and SwedECD dataset pair.

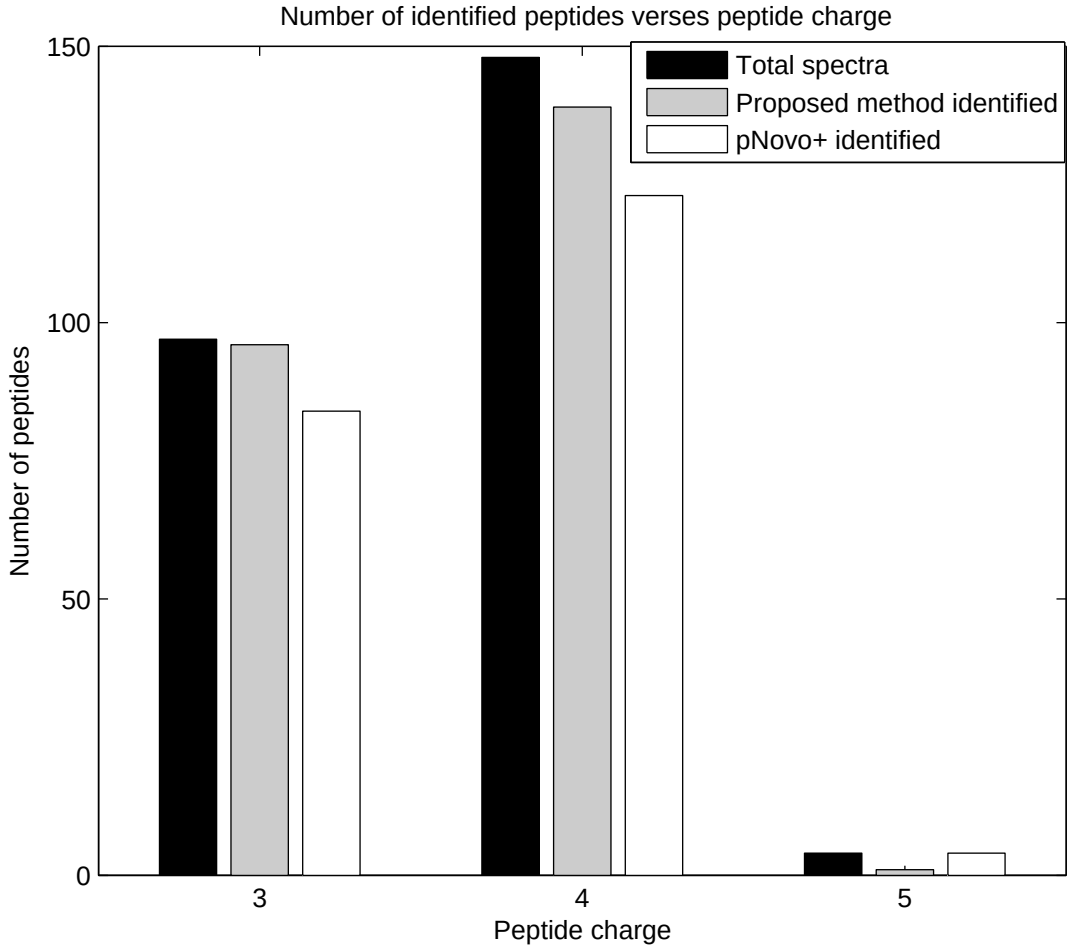


Figure 5.3: Comparison of the number of correctly identified peptides verses peptide charges for the proposed method and pNovo+ on the SCX_HCD_no_decon and SCX_ETD_no_decon dataset pair.

pair in this figure.

Figure 5.3 shows that NovoGMET outperforms pNovo+ for almost all spectra. For the two datasets consisting primarily of +3 and +4 charged spectra, the proposed method identifies more peptides than pNovo+; but on charges +5, the proposed method shows a decrease. The reason for this might be that the complexity of charge +5 spectra causes some problems for the proposed method. However, more experiments are needed to come up with a firm conclusion about this issue.

5.4 Conclusions and future work

In this paper, a new solution to the *de novo* peptide sequencing problem for multiple spectra has been proposed. The proposed framework defines a way to merge multiple experimental spectra, uses a new spectrum graph model (GMET) and different fragment ion types occurring in them, considers length-three peptide tags

to separate a peptide into small regions, and integrates amino acid composition (AAC) information into the graph model. In addition, the proposed method includes a parent peptide correction step, which is essential in solving the multiple spectra sequencing problem.

The proposed framework is applied to peptide sequencing using a pair of spectra from the same peptide. Three pairs of spectral datasets were used to investigate and compare the performance of the proposed method and another successful *de novo* peptide sequencing method, pNovo+. Experimental results have showed that the proposed method outperforms pNovo+ in terms of full length peptide sequencing accuracy, with the accuracy increasing by up to 13.6%. For long peptides (more than 20) and highly-charged (+5) peptides, both methods show decreases to varying degrees, hence they are still challenging problems in *de novo* peptide sequencing. The proposed method, which breaks peptides into smaller parts, and integrates AACs information and GMET model, provides a promising solution to solve these problems.

In future, we will evaluate the proposed method on more MS/MS datasets and compare it to other available paired spectra sequencing methods. In addition, we are also planning to apply the framework to more spectra – for example, CID, HCD, and ETD triplets – for *de novo* peptide sequencing.

Acknowledgment

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

Erratum

In Figure 5.1, the second step should be "Use S_{all} to generate merged spectrum S_m, and conduct mass correction".

Chapter 6

De novo peptide sequencing using CID and HCD spectra pairs

Prepared as: Yan Yan, Anthony J. Kusalik and Fang-Xiang Wu. "*De novo* peptide sequencing using CID and HCD spectra pairs (unpublished)". This is a manuscript resubmitted to Proteomics after a revision in June, 2015.

In the previous chapter, a framework of *de novo* peptide sequencing for multiple tandem mass spectra is presented. It is then applied to paired CID (or HCD) and ECD (or ETD) spectra. The reason for choosing this kind of spectra pair is the complementary features of them. With the availability of various kinds of spectra, other types of spectra combinations could be used for peptide sequencing as well. For example, CID and HCD spectra.

CID and HCD spectra have similar dominant ions. Specifically, they both produce b-ions and y-ions primarily. Additionally, HCD spectra usually have more abundant ions in the low mass region (typically below 200 Da) than CID spectra; to be specific, they produce immonium ions (IMs) and internal ions. The pairing of these spectra has the potential to identify dominant fragment ions with high confidence, and the use of spectrum specific ions can help with the sequencing, too. However, less attention has been paid to CID and HCD spectra pairs currently. Therefore, the study presented in this chapter is to develop a new method specifically for this kind of spectra pair.

In this chapter, CID and HCD spectra are studied and a new sequencing method designed for them is proposed. The proposed method includes a merging criteria for CID and HCD spectra and a parent mass correction step. Experimental results on several MS/MS spectral datasets show that the proposed method outperforms other single-spectrum-based methods and identifies some new peptides that other single-spectrum-based methods cannot identify.

Abstract

In tandem mass spectrometry (MS/MS) there are several different fragmentation techniques possible including collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), electron capture dissociation (ECD), and electron transfer dissociation (ETD). When using pairs of spectra for *de novo* peptide sequencing, the most popular methods are designed for CID (or HCD) and ECD (or ETD) spectra because of the complementarity between them. Less attention has been paid to the use of CID and HCD spectra pairs. In this study, a new *de novo* peptide sequencing method is proposed for these spectra pairs. This method includes a CID and HCD spectra merging criterion and a parent mass correction step, and modifies our previously proposed algorithm for sequencing the merged spectra. Two pairs of spectral datasets were used to investigate and compare the performance of the proposed method with two other methods designed for single spectrum (HCD or CID) sequencing. Experimental results showed that full length peptide sequencing accuracy increased dramatically by using spectra pairs in the proposed method, with the highest accuracy reaching 81.31%.

6.1 Introduction

Tandem mass spectrometry (MS/MS) is a widely used technology for peptide sequencing. MS/MS measures the mass-to-charge ratio (m/z) of the fragment ions of peptides in compounds, and outputs tandem mass spectra (MS/MS spectra) [48] containing m/z values and intensities of ions. In MS/MS, the most common fragment ions are named a-, b-, c-, x-, y-, and z-ions according to the cleavage sites on a peptide backbone that gives rise to them. Different fragmentation techniques used in MS/MS yield differing dominant fragment ions. Collision-induced dissociation (CID) is the traditional fragmentation technique used in MS/MS and it yields b-ions and y-ions as dominant ions. With the development of new fragmentation techniques, alternative MS/MS spectra appeared in recent years. Higher-energy collisional dissociation (HCD) spectra have similar dominant ions as CID spectra but with more abundant ions in the low mass region (typically below 200 Dalton); specifically, immonium ions (IMs) and internal ions [36]. Electron capture dissociation (ECD) and electron transfer dissociation (ETD) preferentially produce variants of c-ions and z-ions, and occasionally a-ions [31–33]. In addition, all the fragment ions usually lose small molecules such as H_2O and NH_3 in the fragmentation process.

Database searching, peptide tagging and *de novo* sequencing [17] are the most popular methods for peptide sequencing using MS/MS spectra. The success of database searching typically relies on the candidate peptides generated from an existing protein database and the effectiveness of the scoring scheme measuring the similarity between theoretical spectra and experimental spectra [46]. *De novo* sequencing does not need a prior database, and thus has the ability to identify proteins that are not included in current databases, proteins resulting from mutations, proteins with unexpected modifications and so on. The main challenge

in *de novo* sequencing is to extract enough information from experimental spectra (usually accounting for noise and missing data) to infer the correct peptide sequences. Peptide tagging [21, 22] is usually used with database searching or *de novo* sequencing to reduce the scale of computation. With the recent development of high mass accuracy MS/MS and alternative fragmentation techniques, *de novo* sequencing has shown promising developments because of the availability of different types of spectra with more information in them [48]. Therefore, this study focuses on MS/MS *de novo* peptide sequencing.

With the appearance of alternative MS/MS spectra resulting from different fragmentation techniques, novel computational methods have emerged to enhance *de novo* peptide sequencing performance. For example, pNovo [72], which employs a spectrum graph model and combines immonium ions (IM) and internal fragment ion information from HCD spectra, has achieved superior peptide sequencing results. Apart from the algorithms based on single spectrum peptide sequencing, many researchers have tried to use multiple spectra from the same peptide to infer peptide sequences [31, 75, 77, 79]. Multiple spectra peptide sequencing extracts information from all spectra, and thus has the potential to increase the accuracy and practicality of *de novo* sequencing [105]. The use of a pair of spectra from CID (or HCD) and ECD (or ETD) fragmentation is the major focus of active study because of the complementary properties of these spectra pairs. Some successful methods include CompNovo [77], ADEPTS [31], pNovo+ [75], and NovoPair [81]. Recently, a universal *de novo* sequencing tool named UniNovo has been proposed [84]. It is based on a spectrum graph model, can be used for various types of spectra, and achieves satisfying sequencing results on various experiential datasets. At this moment, there is still a lack of methods particularly designed for spectra with similar dominant ions, for instance, CID and HCD spectra pairs.

A straightforward way of utilizing CID and HCD spectra pairs for sequencing is to combine the two spectra into a single one. However, more noise may be introduced by this method. Ion redundancy is another problem since the spectra have the same (or similar) dominant fragment ions. Therefore, in order to better use the information from these spectra, a new *de novo* peptide sequencing method is needed. Here, such a method for CID and HCD spectra pairs is proposed. The proposed method includes a merging criterion for CID and HCD spectra and a parent mass correction step that considers the unique features of this spectra pair, and it contains a modification of our previously proposed algorithm for HCD spectrum [73] to perform *de novo* sequencing on merged spectra. Peptide tags and amino acid compositions are used to reduce the scale of computation in the proposed method. The proposed method is then implemented as MATLAB code.

The reminder of the paper is organized as follows: Section 2 presents the design of the proposed method, Section 3 shows the experimental results and performance analysis, and finally Section 4 concludes the paper and gives some directions for future work.

6.2 Methods overview

In this section, the proposed method is briefly introduced. The detailed methodology is in the Additional Files section. The proposed method first considers unique features from experimental spectra to generate a merged spectrum S_m, and then uses fragment ions in the experimental spectra to conduct a parent mass correction. After that, it uses peptide tags from S_m to break a whole peptide sequence into smaller regions. Assume that $Thres$ is the threshold controlling the size of the regions. If the mass range of a region is larger than $Thres$, more tags are used to further decompose it until all regions are no larger than the predefined threshold. A previously proposed method for HCD spectra, NovoHCD [73], with modifications is then applied for sequencing using S_m. Detailed information about NovoHCD and its modification is in the Additional Files section.

Another important aspect of the method's design is the types of fragment ions considered. Some common fragment ions observed in CID and HCD spectra have been introduced in the previous section. Based on the literature [17, 32, 33, 41, 73], ions listed in Table 6.1 are considered in the proposed method. In Table 6.1, m_{CO} is the mass of a molecule of CO. $\sum(residue\ mass)$ is the mass sum of all amino acids from an end amino acid of a peptide sequence to the amino acid at the current cleavage site. b_m and y_m are the masses of the b-ion and y-ion at the current cleavage site, respectively. In Table 6.1, the first column lists the type of ion, and the second and third columns give two ways to calculate the mass of that type of ion. The second column is the calculation based on previous residues, which gives the relationship of consecutive cleavage sites on a peptide backbone. The third column is the calculation based on other ions (either b- or y- ions) at the same cleavage site.

Table 6.1: Ion types considered in CID and HCD spectra

Ion Type	Mass calculation from residues	Mass calculation from other ions
a	$\sum(residue\ mass) - 26.9871$	$b_m - m_{CO}$
b	$\sum(residue\ mass) + 1.0078$	b_m
x	$\sum(residue\ mass) + 44.9977$	$y_m + m_{CO}$
y	$\sum(residue\ mass) + 19.0814$	y_m

In addition, two other ions frequently observed in HCD spectra, immonium ions (IM) and internal fragment ions, are considered. Internal b- and y- ion with length up to two amino acids are also considered. They provide supplemental information for peptide sequencing as well as candidate ranking. Other useful ions include fragment ions at the end of the peptide backbone like b_1-ions and y_1-ions, and a_2/b_2-ion pairs.

Furthermore, the complementary ion relationships in Equations (1)-(2) hold for ions in Table 6.1. In the equations, m_p is the mass of the parent peptide P, N is the peptide length, and $i \in \{1, 2, \ldots, N\}$.

$$|m_p + 2m_H - (b_i + y_{N-i})| = 0 \tag{6.1}$$

$$|m_p + 2m_H - (a_i + x_{N-i})| = 0 \tag{6.2}$$

In the proposed method, all ions introduced above and small molecule loss (such as H_2O and NH_3) are considered.

After determining the types of ions, the next step in the proposed method is to merge the paired CID and HCD spectra into one spectrum. This is done using an iterative selection process. An initial set S_m^0 is created by collecting three types of ions: ions from both spectra with very small mass difference, complementary ion pairs from both spectra, and all IMs from the HCD spectrum. The mass difference threshold for the selection is denoted as δ_1. Any ion pair with very small mass difference is transformed into a single ion in the merged spectrum, with its m/z value being the average and intensity being the sum of the contributing values, respectively. From S_m^0, successive iterative steps create a series of sets S_m^i, where $i \in [1, t]$. To form S_m^i , the iterative step selects ions from the remainder of the original spectra pair where the mass difference between the candidate ion and an ion in the previously generated set S_m^{i-1} is close to one amino acid (mass difference threshold δ_1). In addition, ion intensity information is considered. To be specific, if I_{max} is the highest intensity value of ions in the spectra pair, then a candidate ion must have intensity value greater than or equal to $\frac{I_{max}}{2}$. The reason of having I_{max} divided by 2 is from past experience, and the users can also choose other suitable values. With larger values, the threshold becomes lower, and more ions including more noise could be selected; with smaller values, the threshold becomes higher, and more noise could be filtered out but some fragment ions may be lost, too. Ions meeting these criteria are added into the set S_m^i. The iterative process terminates when no more ions can be selected. As a post-processing step, the remaining ions in the spectra pair are re-examined. Any ion having intensity larger than the two ions on either side of it is added to S_m^{i+1}. All selected ions in S_m^i from the above steps comprise the final merged spectrum S_m.

Since two MS/MS spectra are used in the proposed method, there are two parent masses. The two masses may be different, though they should be close. Thus, a parent mass correction is performed to determine a single, unique mass value for the subsequent peptide sequencing. The details are presented in the Additional Files section.

The proposed method also modifies our previous proposed *de novo* peptide sequencing method for HCD spectra. A major difference is the inclusion of a spectra merging step. Here, we give a simple example to show how the merging step works. Assume the m/z value, intensity pairs of two experimental spectra are $S_c = \{$**(130,0.3)**, $\underline{(182, 0.2)}$, $\underline{(199, 0.5)}$, *(346,0.7)*, **(433,0.5)**, $\underline{(466, 0.3)}\}$ (representing a CID spectrum) and $S_h = \{$**(60,0.2)**, **(130,0.5)**, **(217,0.4)**, $(231, 0.3)$, *(493,0.5)*$\}$ (representing a HCD spectrum), respectively. The meaning of different fonts and underlining will be explained in the following content. Spectra S_c and S_h are also shown in Figures 6.1 and 6.2. The parent mass for both spectra is $m_p = 648$. Mass difference threshold $\delta_1 = 0.01$Da. In order to simplify the calculation and focus on the principles of the method, we use integer m/z values, assume that all ions are in charge state +1, and do not consider loss of small molecules.

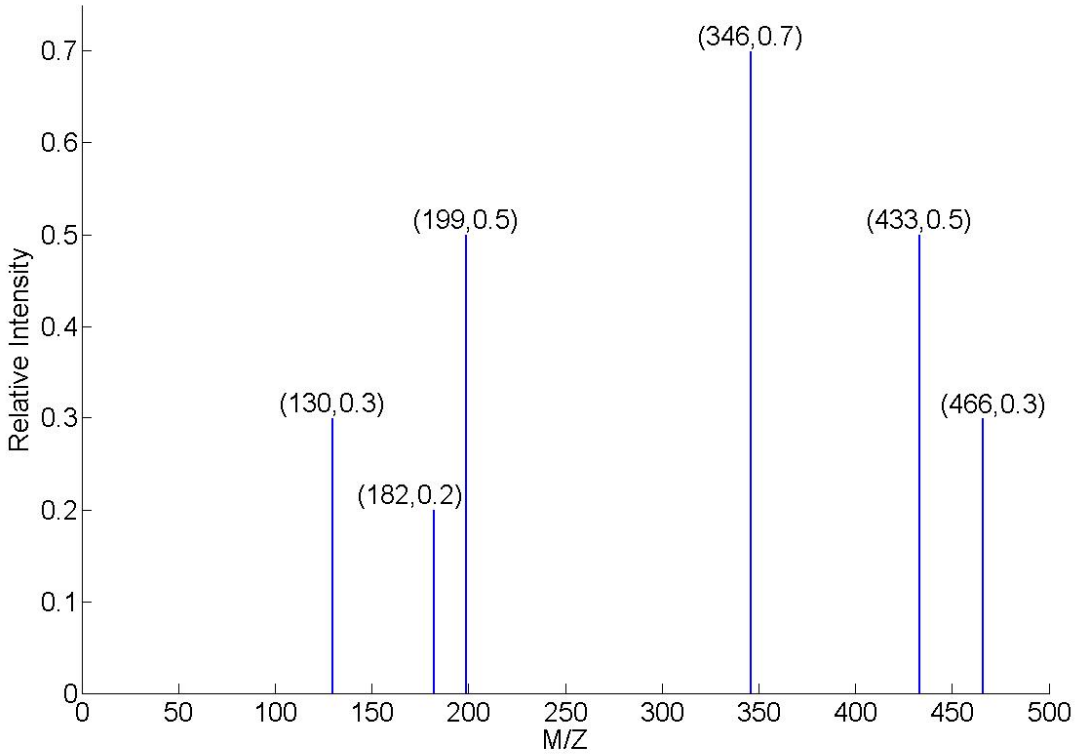


Figure 6.1: The spectrum of S_c

All intensity values in S_c and S_h are in the range of [0,1] to simplify the calculation.

We first find ions forming S_m^0. As shown on spectrum S_c and S_h Figure 6.3 (blue double arrow with label "Same M/Z" on it), since both spectra have m/z value 130, it is added into S_m^0. Its intensity is the sum of two ions' intensity values. During the check for IMs in the HCD spectrum, the first ion (m/z=60) is selected as it indicates glutamic acid (mass difference is within the threshold δ_1). Then we use parent mass to find out complementary ions. Since $\mid 648 + 2m_H - (217 + 433) \mid = 0$, ions having mass values as 217 and 433 are complementary ions. Therefore, $S_m^0 = \{(60, 0.2), (130, 0.8), (217, 0.4), (433, 0.5)\}$ whose elements are highlighted in boldface above and shown in Figure 6.3 as spectrum S_m^0.

Having S_m^0, we conduct the iteration by using amino acid masses and ion intensities. First we consider amino acid mass differences. Pair (130,0.8) is in S_m^0 while the pair (231,0.3) is not yet in the set. We observe that $\mid (231 - 130) - a_T \mid \leq \delta_1$ where a_T is the mass of threonine (referenced ions are highlighted by a black arrow in Figure 6.3). Hence the ion with m/z value of 231 could potentially be added to S_m^1. However, its intensity does not satisfy the intensity criterion. Pair (217,0.4) is in S_m^0 while the pair (346,0.7) is not. We observe that $\mid (346 - 217) - a_E \mid \leq \delta_1$ where a_E is the mass of glutamic acid (referenced ions are highlighted by a black dashed arrow in Figure 6.3). Further the intensity of (346,0.7) satisfies the intensity criterion. Therefore (346,0.7) is added to S_m^1. This is denoted by (346,0.7) appearing in italic in a preceding paragraph.

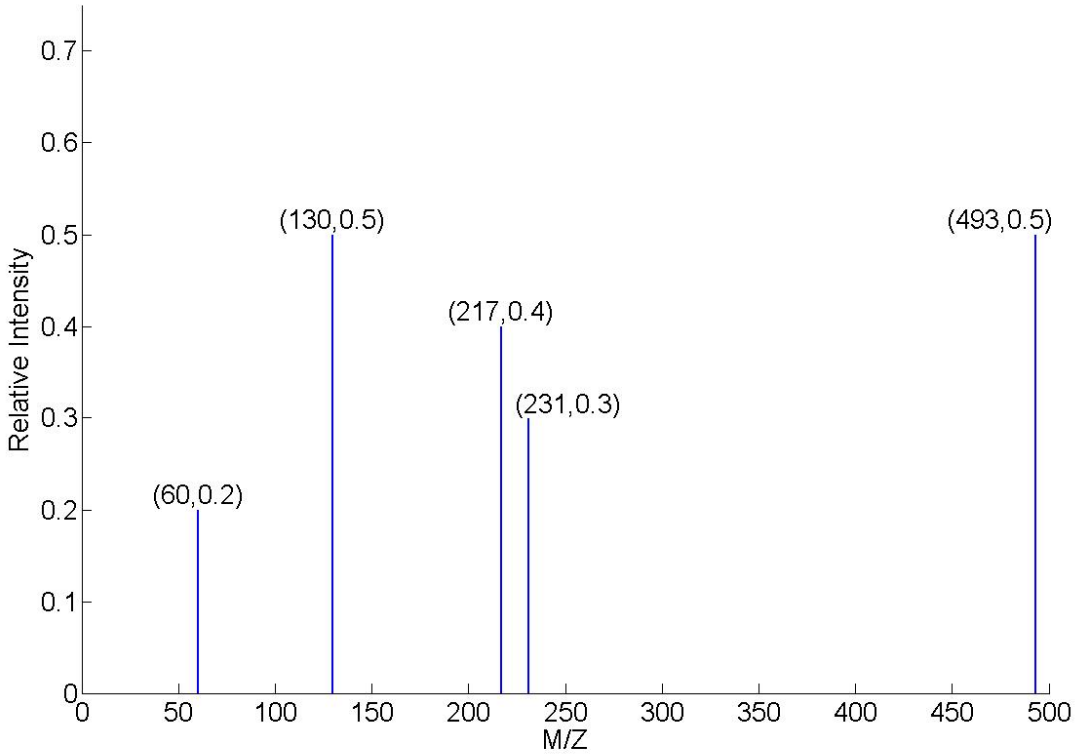


Figure 6.2: The spectrum of S_h

Since there are no more ions in the remaining spectra satisfying the amino acid difference relationship with the ions in S_m^0, $S_m^1 = (346, 0.7)$.

Similarly, pair (346,0.7) is in S_m^1 while the pair (493,0.5) has not been added yet. We observe that $\mid (493 - 346) - a_F \mid \leq \delta_1$ and the intensity of (493,0.5) satisfies the intensity criterion. (493,0.5) is added to S_m^2 (referenced ions are highlighted by a black dashed line in Figure 6.3). This is denoted by (493,0.5) appearing in italic in a preceding paragraph. Since there are no more ions in the remaining spectra satisfying the amino acid difference relationship with the ions in S_m^1, $S_m^2 = \{(493, 0.5)\}$. At this time, as there are no more ions that can be selected by using amino acid masses, the iteration stops. All ions selected from the iteration are added together and shown as spectrum $S_0 \cup S_1 \cup S_2$ in Figure 6.3.

Finally, by checking the intensities of the remaining ions in both spectra, ion (199,0.5) is selected. $S_{m^3} = \{(199, 0.5)\}$. In this example, ions (182,0.2), (199,0.5), and (466,0.3) (underlined above and shown in the spectrum S_c^{left} in Figure 6.3) are left from the CID spectrum and none are left from the HCD spectrum. Therefore, in this example, the final merged spectrum of the HCD and CID spectra pair is $S_m = \{(60, 0.2), (130, 0.8), (217, 0.4), (433, 0.5), (346, 0.7), (493, 0.5), (199, 0.5)\}$, as shown on the spectrum S_m in Figure 6.3.

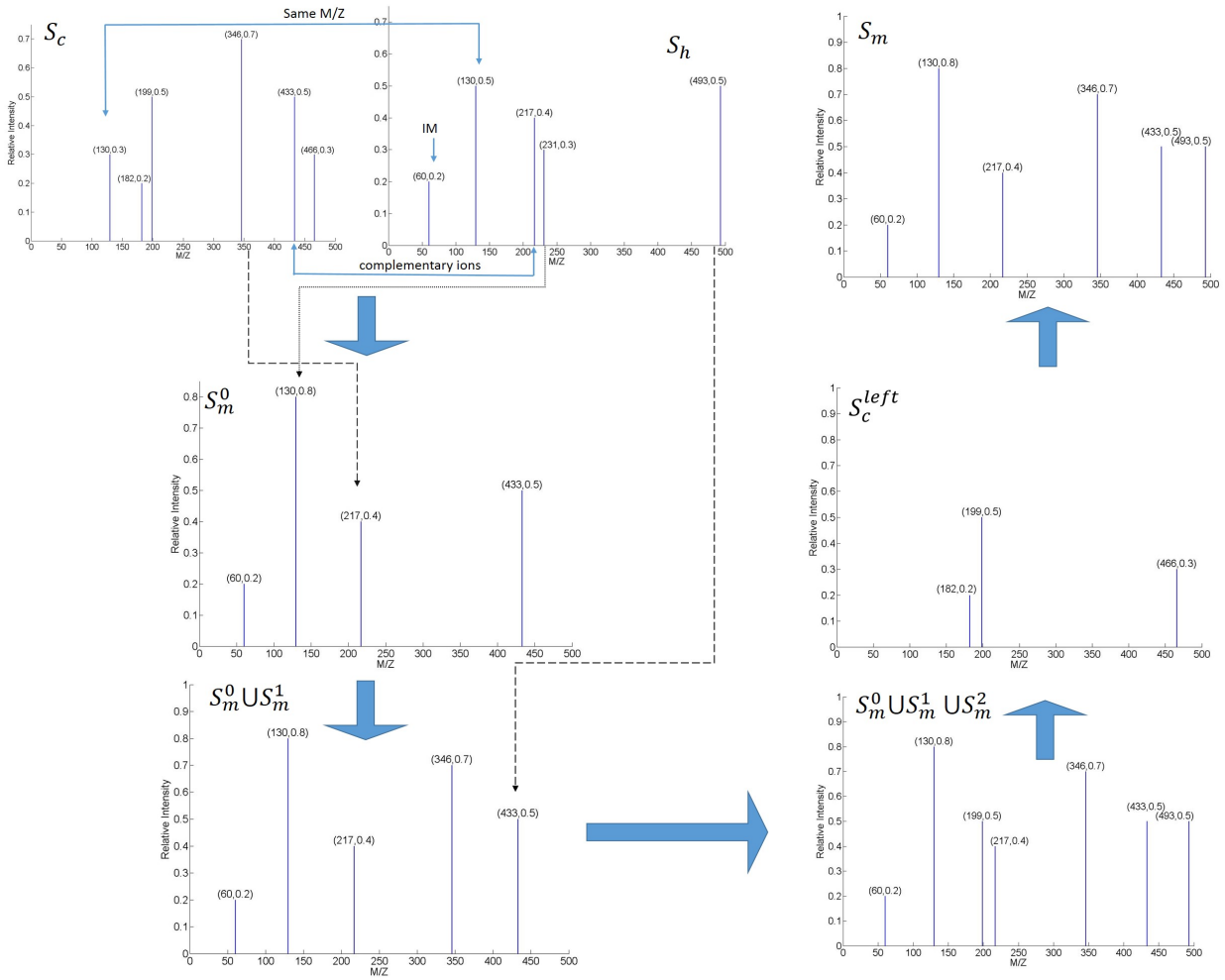


Figure 6.3: An example of the spectra merging process

6.3 Experiments and Results

In this section, the performance of the proposed method for CID and HCD spectra pairs is evaluated. Our previously proposed *de novo* peptide sequencing method for HCD spectra, NovoHCD [73], and another method for HCD and ETD spectra, pNovo+ [75] (with the option for HCD spectra), are used for comparison. More information about NovoHCD and pNovo+ can be found in the Additional Files section. We also tried to use UniNovo [84] in the comparison. However, instead of reporting whole peptide sequences, UniNovo produced partial peptide sequences along with numerical values representing the mass sums of unidentified parts. Hence, UniNovo is not included in the comparison. The reasons for choosing NovoHCD and pNovo+ are as follows. First, the proposed method focuses on the merger of the two experimental spectra, and uses a sequencing method similar to the one in NovoHCD. The result difference between the proposed method and NovoHCD is expected to show the contribution of using merged spectra. Secondly, to our knowledge, there is no more method available for CID and HCD spectra pairs other than the proposed one. Considering the

similarity between the two types of spectra, a comparison between the proposed method and other methods designed for HCD spectra is expected to show the improvement of using spectra pairs. Lastly, pNovo+ is used to conduct peptide sequencing on the spectra generated by the spectra merging step in the proposed method. A comparison of pNovo+ and proposed method in this case can show the effectiveness difference of peptide sequencing models in these two methods. The detailed experimental process and results are presented below.

6.3.1 Datasets

There are two pairs of CID and HCD spectral datasets used in the experiments: SCX_CID_decon and SCX_HCD_decon, plus SCX_CID_no_decon and SCX_HCD_no_decon [98]. Both dataset pairs are from the data used in [98]. MS/MS spectra were generated from a hybrid tandem mass spectrometer; they were analyzed on an ETD enabled Orbitrap Velos instrument (Thermo Fisher Scientific, Bremen) connected to an Agilent 1200 HPLC system [98]. The experimental spectra were then interpreted using Mascot software version 2.3.02 (Matrix Science, UK). Details about the sample, instrument, and parameters used for generating the expermental spectra can be seen in [98].

The original datasets contain various fragmentation MS/MS spectra including CID, HCD, and ETD spectra. The CID and HCD spectra were chosen for the experiments here. The latter dataset pair (labelled with "_no_decon") contains spectra without deconvolution while the other pair contains spectra with deconvolution. Each spectrum comes with a precursor ion charge value and a peptide sequence indicating the peptide as determined by Mascot. In order to conduct our experiments, spectra pairs having the same peptide sequence were selected from the paired datasets.

When each spectrum in a spectra pair is used separately and single-spectrum-based methods output identical and correct results, there is no need to use the proposed method for paired spectra sequencing. Such spectra pairs are of sufficient quality that any one of the two produces satisfying results. The remaining paired spectra constitute a much more rigorous test and those were used in testing the proposed method and comparing it with the other sequencing algorithms. We used NovoHCD as the single-spectrum-based method to implement this selection strategy for constructing the test data set. The spectra charges and the numbers of selected pairs of spectra in the testing datasets are summarized in Table 6.2.

Table 6.2: Number of selected spectra pairs and charges in each dataset used in the experiments

Dataset pair	Charge of selected spectra	Number of selected spectra
SCX_CID_decon SCX_HCD_decon	+2 to +6	403
SCX_CID_no_decon SCX_HCD_no_decon	+2 to +5	578

Table 6.3: Parameters used in the experiments

Parameter	Role in proposed method	Value
δ_1	Generation of S_m^0 in spectra merging	0.01Da
δ_2	Iteration in spectra merging	0.01Da
$Thres$	Multiple tags used in modified NovoHCD	1600Da
Number of tags	Tag integration in modified NovoHCD	10 per spectrum
Number of output sequences	Candidate output	3 per spectrum

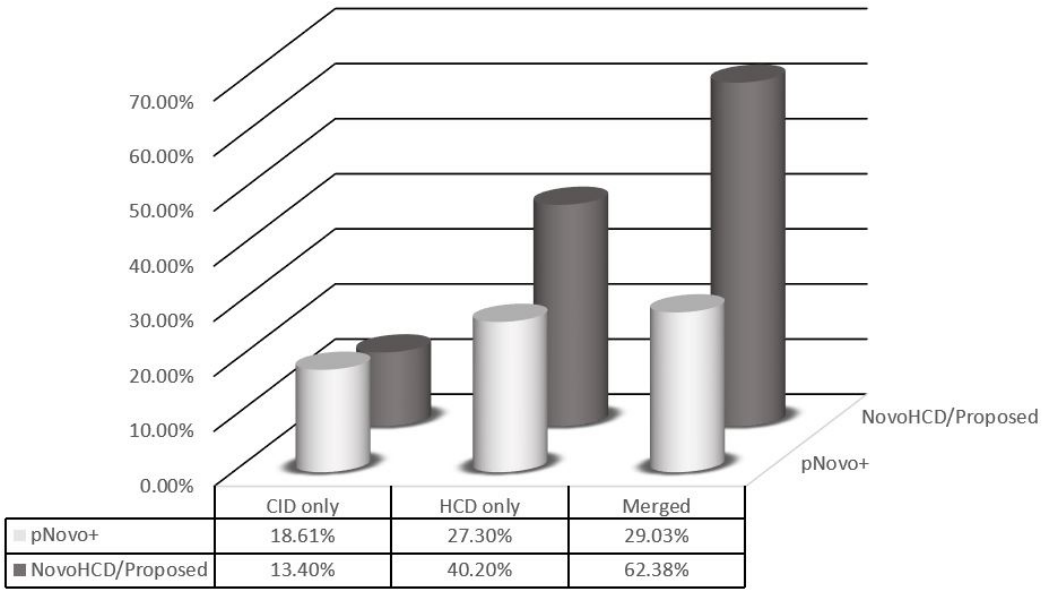
6.3.2 Parameters

There are several parameters in the proposed method, and the values used are listed in Table 6.3. All these values are set according to our previous study and experiments [73, 100], but they can be changed by users to suit their needs.

6.3.3 *De novo* peptide sequencing performance

We investigated the performance of different methods by comparing full length peptide sequencing accuracy. For each spectrum (merged or experimental), a series of candidate sequences are output with ranking scores associated with them. A higher rank indicates greater confidence in the correctness of the predicted sequence. The top three ranked candidates are considered in the performance comparison. If any one of the three candidates interpreted from a spectrum is correct, we say that the method achieves a full length accuracy for the given spectrum. The results for full length accuracy comparison are presented in Figures 6.4 and 6.5. In the comparison, results from NovoHCD and the proposed method are categorized together and compared with pNovo+. One reason for that is that the models of NovoHCD and the proposed method are similar, so they can be categorized together and compared to pNovo+ (HCD spectra option). The other reason is that NovoHCD is applied to each of HCD and CID spectra alone, the proposed method is applied on merged spectra only, and pNovo+ is applied on both cases. Therefore, NovoHCD and the proposed method can be categorized together.

From Figures 6.4 and 6.5 one can see that for both dataset pairs, the proposed method achieves higher full length accuracy than NovoHCD when using merged spectra. Since the two methods have similar sequencing models, this comparison implies that the merger of the two spectra successfully provides additional information for better peptide sequencing without introducing excessive noise or losing essential information. The proposed method also outperforms pNovo+ where the latter uses CID or HCD spectra alone. This proves the significance for peptide sequencing by using paired spectra with similar dominant ions. When pNovo+ conducted peptide sequencing on the spectra generated from the spectra merging step in the pro-



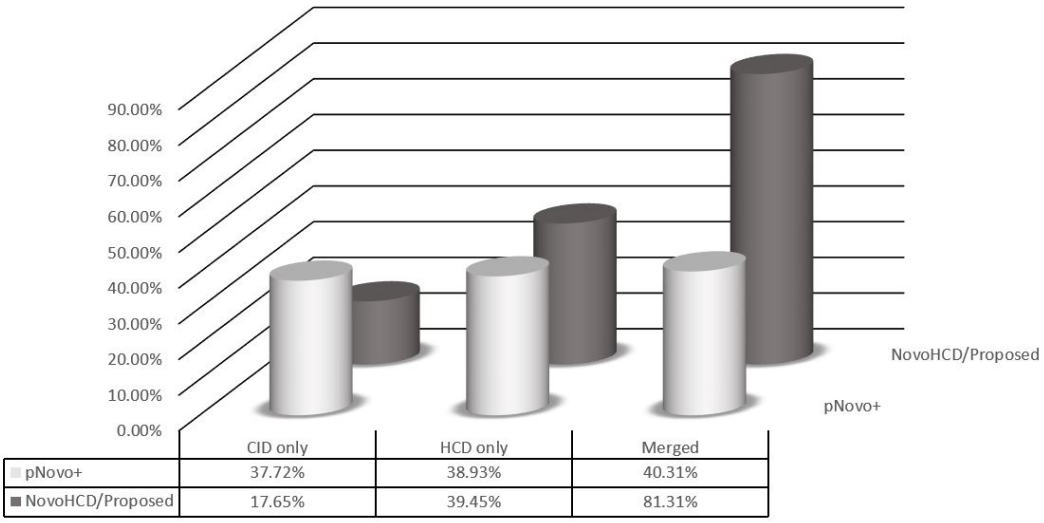
	CID only	HCD only	Merged
pNovo+	18.61%	27.30%	29.03%
NovoHCD/Proposed	13.40%	40.20%	62.38%

Figure 6.4: Full length sequencing accuracy comparison on SCX_CID_decon and SCX_HCD_decon testing datasets.

posed method, it achieved a lower number of correctly identified peptides on both dataset pairs. This implies an improvement of the sequencing model in the proposed method as compared to pNovo+. Finally, it is noticeable that pNovo+ does not have a dramatic difference on these three cases while the proposed method along with NovoHCD does have a steady performance improvement. This may be because pNovo+ is designed for HCD spectra and/or ETD spectra but not HCD and CID spectra pairs, and the model may not work better for the merged spectra with additional information. In contrast, the proposed method, with an extended model based on NovoHCD, is designed for merged spectra. The proposed method successfully takes use of additional information from merged spectra and thus has better accuracy than pNovo+, even when the latter is given the same spectra generated the merging step in the proposed method.

One key feature contributing to the success of the proposed method is the spectra merging step. This step extracts additional signal ions from spectra pairs without introducing excessive noise, which provides a sound foundation for the graph model built upon the merged spectra. In the algorithm of pNovo+, one ion in a spectrum is converted into several vertices in the graph model, which could introduce additional noise. This may effect its performance on the experimental datasets. The other key feature of the proposed method is the model for sequencing. Multiple types of edges in the graph model represent different relationships between vertices, and the AACs limit the numbers of edges and thus reduce the incidence of false positive amino acids in the predicted candidate sequences. With the use of multiple tags, the number of possible AACs can be restricted to an acceptable scale ($\leq$ 1600Da), and the amount of calculation is reduced.

When using paired spectra for peptide sequencing, since two spectra are used, additional information for the task can be extracted. When using the proposed method, the number of identified peptides using spectra



	CID only	HCD only	Merged
pNovo+	37.72%	38.93%	40.31%
NovoHCD/Proposed	17.65%	39.45%	81.31%

Figure 6.5: Full length sequencing accuracy comparison on SCX_CID_no_decon and SCX_HCD_no_decon testing datasets.

pairs are larger than the total number of peptides identified by using CID and HCD spectra separately. One can arrive at this conclusion by transforming the accuracies in Figures 6.4 and 6.5 into total numbers of identified peptides given the total number of experimental spectra. This suggests that for some spectra pairs, individual CID and HCD spectra may contain partial information for sequencing, but not enough for the whole peptide sequence to be identified. Then, with the combination of the two, this information can be combined, and it becomes sufficient for sequencing. The spectra pairs for which NovoHCD failed to output correct sequences when using either CID or HCD spectra alone were noted. The numbers of these spectra pairs from the two experimental dataset pairs are shown in Table 6.4 under the heading "number of failed pairs". The numbers of these spectra pairs successfully sequenced using the proposed method are shown in the last column of Table 6.4. From Table 6.4 one can see that there are many spectra pairs in each dataset pair for which neither CID nor HCD spectra alone are sufficient for successful sequencing. However, correct sequences of many of them are identified by the proposed method when using both spectra. This shows that the proposed method for spectra pairs can fill gaps in the capabilities of single-spectrum-based methods for *de novo* peptide sequencing. If a method processes the two spectra separately and combines the sequencing results later, it may not be able to identify all of the new peptides as the proposed one does. Combining the informative ions from both spectra makes it easy to do a successful the sequencing based on graph-theoretic algorithms since there are sufficient information in the merged spectrum.

Furthermore, we considered the relationship between the number of correctly identified peptides and peptide length. Comparisons between the proposed method and pNovo+ on dataset pair SCX_HCD_no_decon and SCX_ETD_no_decon are shown in Figures 6.6, 6.7 and 6.8. The reason for showing the results from this pair is because it has a larger number of spectra, and so is expected to constitute a more comprehensive

Table 6.4: Number of successful sequencing pairs from the ones for which sequencing failed by using CID and HCD spectra alone

Dataset pair	Number of failed pairs	Number of peptides successfully sequenced by proposed method
SCX_CID_decon SCX_HCD_decon	187	80
SCX_CID_no_decon SCX_HCD_no_decon	248	181

comparison.

Figures 6.6, 6.7 and 6.8 show that the proposed method outperforms pNovo+ on every peptide length in this pair of datasets. The identification accuracy of pNovo+ drops after length 16. However, the proposed method achieves very good results when peptide length is greater than 16, but less than 20. (Beyond 20, the proposed method shows some decrease.) This indicates that use of multiple tags in the proposed method contributes to the sequencing of long peptides. Our previous study on ETD and ECD spectra showed that with the use of multiple spectra, peptide sequencing accuracy can be greatly improved on long peptides (typically $\geq$ 15 amino acid long), compared to single tag usage [100]. pNovo+, however, does not include a strategy designed specifically for long peptides, and this may be a reason for the unsatisfying performance on these peptides. The varied performance of the proposed method on very long peptides ($\geq$ 20 amino acid long) may reflect the complexity of the spectra generated by these peptides. Such complexity could cause the drop of the performance. However, since the number of the spectra generated by these very long peptides is limited in the experimental datasets, we believe that future experiments and analysis are needed to investigate this phenomenon.

6.4 Conclusions and future work

In this paper, a new method of *de novo* peptide sequencing for CID and HCD spectra pairs is proposed. The proposed method includes a criterion for merging pairs of CID and HCD spectra and a parent mass correction technique. The method is a modification of our previously proposed NovoHCD method. Peptide tags and amino acid compositions are used to reduce the scale of computation in the proposed method.

Two pairs of spectral datasets were used to investigate and compare the performance of the proposed method and two other methods designed for single spectrum (HCD or CID) data, NovoHCD and pNovo+ (HCD option). Experimental results showed that full length peptide sequencing accuracy increased dramatically through the use of spectra pairs in the proposed method, with the highest accuracy of 81.31%. The

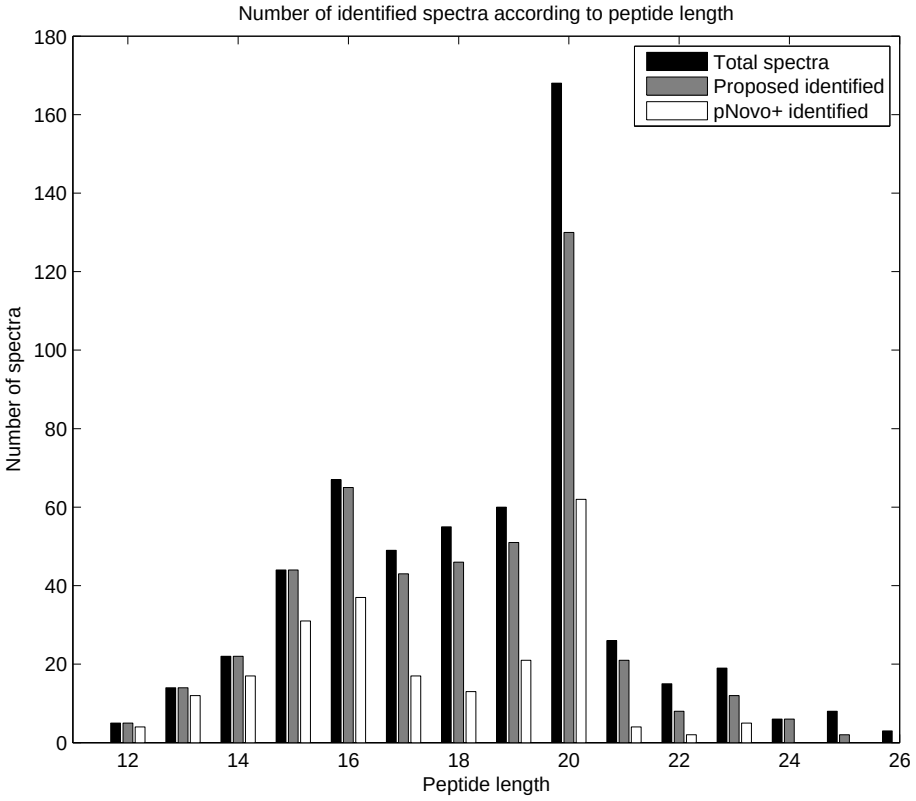


Figure 6.6: Comparison of the number of correctly identified peptides verses peptide length for the proposed method and pNovo+ using only CID spectra for the latter method.

proposed method has better peptide identification accuracy than either single-spectrum-based method.

In future, we will focus on improving the accuracy of the proposed method by further analyzing the characteristics of the wrongly identified spectra, and evaluating the proposed method on more MS/MS datasets. In addition, we are also planning to design a universal framework of *de novo* sequencing for all kinds of MS/MS spectra including CID, HCD, ETD spectra and their combinations.

Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

Competing interests

The authors declare that they have no competing interests.

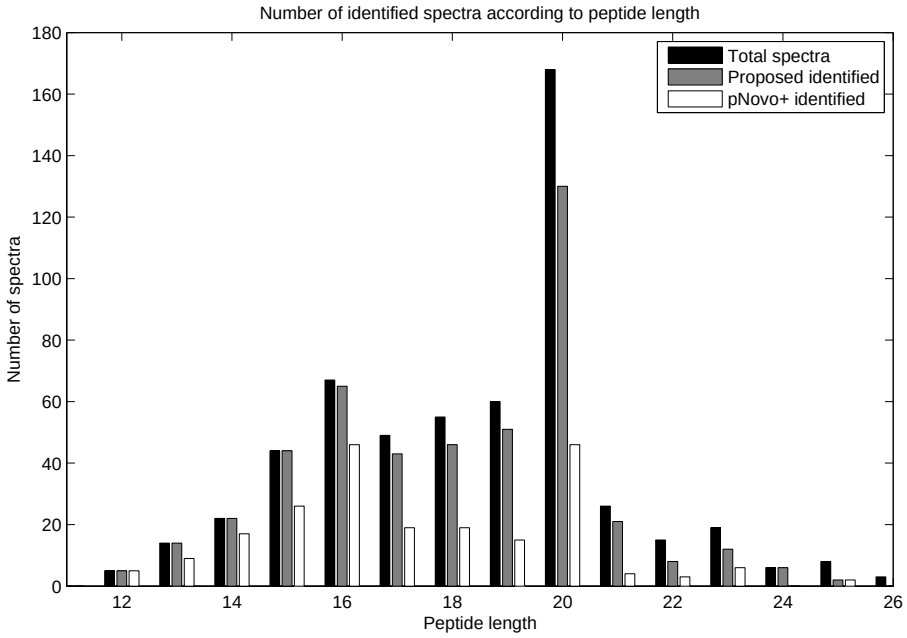


Figure 6.7: Comparison of the number of correctly identified peptides verses peptide length for the proposed method and pNovo+ using only HCD spectra for the latter method.

Additional Files

This section describes the proposed method in detail including the spectra merging and parent mass correction steps, and the how it extends NovoHCD [73].

Spectra merging

Since CID and HCD have similar dominant ions, the same fragment ions may occur in both spectra; for example, a b_3-ion observed in both experimental spectra. Therefore, in the merging step, such ions need to be combined. In the proposed merging criterion, ions from both spectra with very small mass difference are first selected. These two ions are more likely to be real fragment ions than noise since the same noise peak is unlikely to appear in both spectra. The two ions are then transformed into one ion in the merged spectrum with its m/z value being the average and intensity being the sum of the individual values from the two ions, respectively. The second feature considered as a criterion in the spectra merging is complementary ion pairs [90]. A pair of ions can be from the same spectrum or one from each spectrum; for example, a b-ion from a CID spectrum and its complementary y-ion from a HCD spectrum. Such ions are added into the merged spectrum. IMs and internal ions introduced previously are selected from the HCD spectrum and also added to the merged spectrum. We denote the ions selected (or generated) from the above steps as S_m^0. The mass difference threshold used in generating S_m^0 is denoted as δ_1.

Having S_m^0 as an initial set, an iterative approach is used to select more ions to add to the merged spectra.

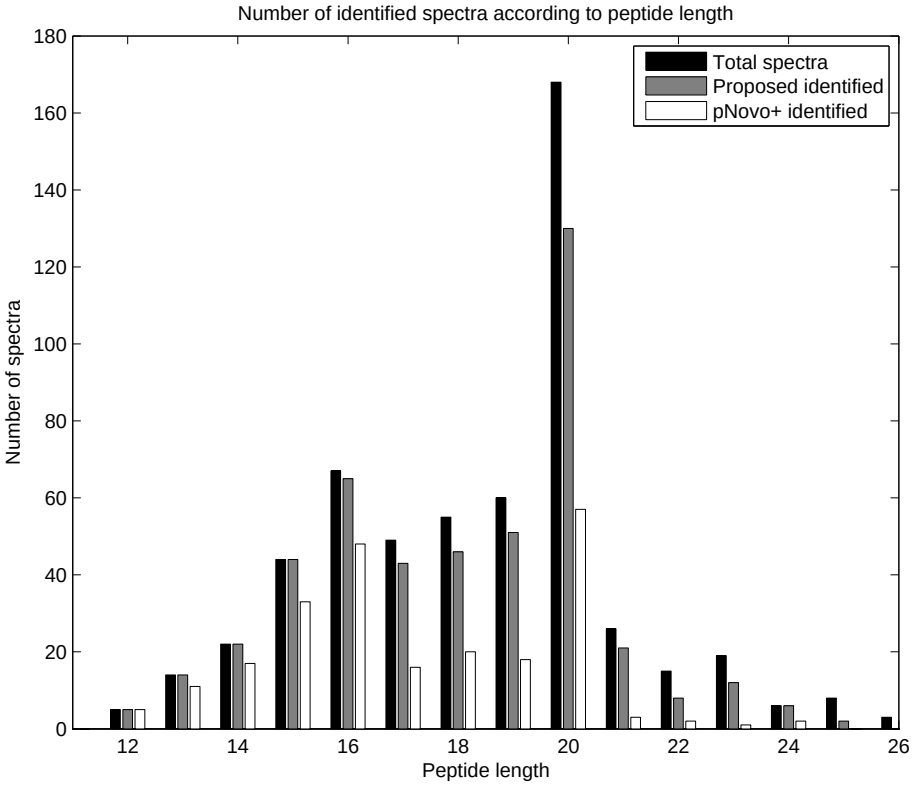


Figure 6.8: Comparison of the number of correctly identified peptides verses peptide length for the proposed method and pNovo+ using merged spectra.

Here, amino acid difference is applied to select additional ions from CID and HCD spectra pairs. For an ion $u \in S_m^0$, if there is an ion v in the CID or HCD spectrum such that the m/z difference between u and v is close to any of the 20 amino acid masses (within a given threshold δ_2), v is selected. We denote all ions selected from this step as S_m^1. Having S_m^1, we use the amino acid difference to select additional ions from the rest of the spectra pair that have one amino acid difference to the ions in S_m^1, and denote the ions selected as S_m^2. This step continues iteratively until no more ions can be selected using the amino acid difference measurement. A suitably designed threshold can be used as the stop criterion for the iteration. The last ion set selected is denoted as S_m^t. One potential benefit of the iteration is that we can assign weights to the ions selected from different rounds of iteration. The hypothesis is that ions in the initial set S_m^0 have the highest confidence to be real fragment ions rather than noise, and as the iteration continues, the level of confidence drops. In the current method, we did not introduce weights in order to simplify the calculation.

Another unique aspect of the proposed method is that it handles high confidence ions (similar m/z, IMs, and complimentary ions) and other ions (selected using amino acid masses) separately. The former does not need additional selection criteria, while the latter can benefit from them because of the possible false positives introduced by comparing all 20 amino acid masses. The proposed method uses ion intensities as an additional

selection control criterion. To be specific, if we denote the highest intensity value of the ions in the spectra pair as I_{max}, ions having intensity values less than $\frac{I_{max}}{2}$ are not selected. The reason behind this criterion is that usually ions with lower intensity are more likely to be noise rather than fragment ions. An intensity comparison (low or high) is only meaningful within one spectrum. In order to use an intensity criterion on a pair of spectra, intensity normalization is needed. The highest intensity values in both spectrum are set to 1, and the remaining intensity values are normalized according to the ratio to the highest ones. After this process, the two spectra have the same scale of intensity values, and the intensity criterion can be applied. Amino acid composition information is used in a subsequent stage of the method to further reduce false positives.

In the last step of the spectra merging process, the remaining ions in the spectra pair are examined. The "local maximum" ions [88] are put into set S_m^{t+1} in order to balance the effect of the above intensity threshold $\frac{I_{max}}{2}$. A peak is called a "local maximum" if its intensity is larger than the two peaks beside it. Researchers have found that when applying intensity information, a simple threshold is not completely effective for differentiating signal ions from noise ions because the ions' intensities in a spectrum tend to be larger in the middle of the m/z range than at the two ends for CID and HCD spectra [88]. It is more reasonable to assume that the noise ions in a narrow m/z range are equally distributed, and that signal ions tend to be the local maxima [88]. Finally, the merged spectrum $S_m = \bigcup_{i=0}^{t+1} S_m^i$.

Parent mass correction

In this method, two MS/MS spectra are used and each comes with a parent peptide mass. Before the *de novo* sequencing, parent mass correction is conducted. The approach introduced in [81] is used here. In this approach, all complementary ion pairs in S_m are used, and the hypothesis is that the real parent mass has the minimal mass difference to complementary ion pair masses. Then by solving the following quadratic expression (6.3), we get the optimal parent mass.

$$\begin{aligned} &\min \Sigma_{j=1}^{k} \mid (I_j + I_j^c) - P_{mass} \mid^2 \\ &s.t. \;\; P_{inf} \leq P_{mass} \leq P^{sup} \end{aligned} \tag{6.3}$$

In expression 6.3, I_j, I_j^c are complementary ion pairs in S_m, k is the total number of complementary ion pairs, and P_{mass} is the optimal parent mass. P_{inf} and P^{sup} are the infimum and supremum of P_{mass}, respectively, which can be set as the masses of the two experimental spectra.

NovoHCD and its modification

NovoHCD [73] is a recent solution to the *de novo* peptide sequencing problem for HCD spectra. It is based on multi-edge graphs with integration of amino acid composition (AAC) and peptide tags.

NovoHCD first uses peptide tags to separate a whole peptide sequence into three parts: prefix, tag, and suffix. Several tags are used for one spectrum and each tag is used separately. For each tag, NovoHCD

builds a multi-edge graph model on the prefix and suffix separately to find partial peptide sequences, and AAC information is used to limit the number of edges in the graph. NovoHCD finally combines the three parts, each possible prefix and suffix sequences and the tags, to generate a candidate peptide. All possible candidates based on different tags are ranked through a ranking scheme and output.

A multi-edge graph G includes five different types of edges reflecting the relationships of complementary ions, amino acid difference between ions, and loss of small molecules of ions. Detailed definition of edges can be seen in [73]. Having graph G, another kind of graph, induced by all vertices in G from a single cleavage site of the peptide, can be constructed. This is named as a "basic structure" of G at a cleavage site. Then, by finding adjacent basic structures, continuous amino acids can be inferred. The detailed steps of inferring basic structure (and amino acids) can be seen in [73].

pNovo and its comparison with NovoHCD

pNovo [72], an alternative software designed for HCD spectra, is used for performance comparison with NovoHCD in our previous study [73]. pNovo applies a graph-theoretic approach for sequencing. It first converts each ion in a spectrum into several vertices representing different types of ions and merges vertices with similar masses, and then connects two vertices if their mass difference is close to one of the 20 amino acid masses. After that, it assigns weights to edges and generates best score paths from the graph model using a scoring function designed by the authors.

Differences between NovoHCD and pNovo are in the graph model and scoring scheme. pNovo converts ions from an experimental spectrum into several vertices, which could introduce additional noise into the graph since one noisy ion is converted into several noisy vertices in the graph. With additional noisy vertices, false positive edges can be added into the graph when comparing mass difference between two ions. This could make the path finding very difficult and result in output of incorrect paths. In contrast, NovoHCD uses multiple types of edges to represent different types of relationships between two ions, and no added vertex is introduced. When forming edges of the graph in NovoHCD, AAC is used to reduce false positive edges. The scoring scheme in pNovo includes a weight assignment and path scoring function, while NovoHCD uses a ranking-based scoring scheme for peptide candidates. Experimental results on several HCD datasets show that NovoHCD outperforms pNovo [72] in terms of full length sequencing accuracy.

NovoHCD must be modified for use in the proposed method. The major modifications are to the peptide tag strategy and the candidate ranking scheme. In NovoHCD, a single tag is used for HCD spectra to separate whole sequences into three parts. Upon further study, we found that this strategy is not effective for long peptides (typically over 15 amino acid long). Therefore, multiple tags are used, as in [17], to separate the whole sequences, and a predefined threshold (set to be 1600 Da for the proposed method) is used to control if further separation is needed.

For the ranking scheme, NovoHCD considers parent peptide mass P_{mass} and the number of IMs in a HCD spectrum. The magnitude of mass difference between a peptide candidate sequence P_{can} and P_{mass} is

denoted as Δm_P. In the proposed method, internal ions in the merged spectrum are also included. Internal ions in the form of *b*- and *y*- ions with length of up to two amino acids are considered. Given P_{can}, we list all internal ions from it and compare them to the merged spectrum. The number of matched internal ions and the number of IMs are added, and denoted as AA. For a given pair of spectra, the ranking score of P_{can} can be represented as a vector $CP(AA, \Delta m_P)$. The ranking approach first sorts CP according to its initial element AA in decreasing order, and then arranges Δm_P in increasing order as a secondary key. After this process, all candidates are ranked. In the proposed method, for a pair of input spectra, the top 3 ranked candidates are output.

Chapter 7

Summary and Future Work

7.1 Summary

Peptide sequencing from MS/MS has become an important topic in proteomics. It provides essential information for protein structure and function study. With the development of MS/MS ionization techniques and various MS/MS spectra generated from them, suitable computational methods are needed for the data interpretation. Currently, *de novo* peptide sequencing methods have drawn a lot of attention because of their unique advantages compared to database searching. Many algorithms for *de novo* peptide sequencing have been developed with application to different types of MS/MS spectra. The main limitations of current *de novo* peptide sequencing methods are the lack of suitable models reflecting MS/MS spectra, limited information extracted from the spectra, and inefficient use of multiple spectra. This thesis aims to address some of the limitations in current peptide sequencing methods with the four objectives listed in Chapter 1. The work presented in Chapters 2 to 6 of the thesis has achieved these objectives.

Chapter 2 presents a comprehensive review of *de novo* peptide sequencing methods and achieves Objective 1. It summarizes recent developments of computational methods for various types of typical experimental data, compares and analyzes their advantages and disadvantages, points out the limitations of current studies, and identifies directions for improvements and new method design.

In Chapter 3, a new model containing useful information from the MS/MS spectra is developed for *de novo* peptide sequencing. It modifies the traditional spectrum graph model to be a graph with multiple types of edges and integrated amino acid combination (AAC) information and peptide tags. This method is then evaluated on HCD spectra and compared with another competing method on several experimental datasets. Results show that it outperforms other methods over the five datasets.

Based on the success of the method proposed in Chapter 3, other types of MS/MS spectra, ECD and ETD spectra, to be specific, are studied in Chapter 4. A *de novo* sequencing method named NovoExD designed for ECD and ETD spectra is presented in this chapter. NovoExD modifies the previous model in Chapter 3 and considers multiple peptide tags and fragment ion charge information. Experiments conducted on three different datasets show that NovoExD outperforms another similar method. Objectives 2 and 3 have been achieved by the methods developed in Chapters 2 and 3.

Objective 4 is about methods for multiple spectra. Chapter 5 presents a framework for multiple spectra

sequencing with new features and models. It is applied to paired CID (or HCD) and ECD (or ETD) spectra. These spectra pairs have different dominant fragment ions and are complementary to each other. There are already some other methods for these spectra pairs, and results on several experimental datasets show that the proposed method outperforms similar methods available.

Chapter 6 presents a *de novo* peptide sequencing method for CID and HCD spectra pairs. These spectra pairs have similar dominant ions but are accompanied by other ion types with different properties. Less attention has been paid in the literature to these spectra pairs. Experimental results show that the proposed method works well on several testing datasets, and identifies some new peptides that other single-spectrum-based methods cannot identify. Objective 4 is accomplished by these two chapters. Therefore, all objectives proposed for the thesis have been achieved.

To sum up, the following works have been completed in this thesis:

- Reviewed *de novo* peptide sequencing methods, analyzed their advantages and disadvantages, and found limitations and potential improvement directions.

- Based on the literature review, proposed a new graph model for *de novo* peptide sequencing and applied it to HCD spectra.

- Revised the proposed method for HCD spectra, and developed a new *de novo* peptide sequencing method for ECD and ETD spectra by considering their unique features.

- Developed a framework for *de novo* peptide sequencing of multiple spectra and applied it to paired CID (or HCD) and ECD (or ETD) spectra.

- Proposed a new *de novo* peptide sequencing method for CID and HCD spectra pairs including a specifically designed spectra merging criteria and modification of a previously proposed method for paired spectra sequencing.

7.2 Contributions

The thesis provides a series of novel computational *de novo* peptide sequencing methods for different types of MS/MS spectra. All proposed methods have been evaluated on several experimental datasets and compared with other methods. In the following, specific contributions are listed:

- A comprehensive literature review of *de novo* peptide sequencing methods of MS/MS spectra is given. It summarizes the development of the methods, analyzes the advantages and disadvantages of different methods, and provides guidelines for filling gaps between current methods and developing advanced methods as further study.

- An improved graph model for *de novo* peptide sequencing is proposed with multiple types of edges reflecting different relationships between ions in an MS/MS spectrum. It is applied to a new type of MS/MS spectra, HCD spectra, for peptide sequencing.

- A new *de novo* peptide sequencing method is proposed with multiple peptide tags and amino acid composition information included to reduce the complexity and increase accuracy of the method. It is applied to ECD and ETD spectra with the consideration of their unique features.

- A framework for *de novo* peptide sequencing of multiple spectra is developed with application to paired CID (or HCD) and ECD (or ETD) spectra. The framework provides general guidelines for the *de novo* peptide sequencing of multiple spectra, and is potentially applicable to various spectra combinations with suitable modification.

- A new *de novo* peptide sequencing method for CID and HCD spectra pairs is developed. It includes a specifically designed spectra merging criteria and modifies a previously proposed method to suit CID and HCD spectra pairs.

7.3 Future Work

Based on the work presented in the thesis of *de novo* peptide sequencing, the following directions for future research work are proposed.

- *De novo* peptide sequencing methods for spectra with post-translational modifications (PTMs).
  One major advantage of *de novo* peptide sequencing is its independence of a protein database. Thus it has the ability to identify peptides with PTMs. Among all kinds of spectra, ECD and ETD spectra are the most suitable for identifying peptides with PTMs because of the unique features of these spectra. A computational method for ECD and ETD spectra without consideration of PTMs has been presented in this thesis (NovoExD). With suitable modification, it can be extended to the case with PTMs. A practical way to start is to consider one common PTM, and make adjustment of the amino acid masses used in the method. After that, more PTMs could be considered. One recent study has claimed that only a limited number of PTMs can occur in a peptide although the types of PTMs can be numerous [106]. Therefore, researchers should consider limiting the number of PTMs per spectrum in the designed method.

- New methods for multiple spectra *de novo* peptide sequencing.
  Multiple spectra sequencing continues to increase in popularity with the availability of various kinds of MS/MS spectra. Two computational methods have been presented in the thesis for different pairs of spectra. However, there is still room for new for multiple spectra sequencing methods for other spectra combinations; for example, CID, HCD, and ETD.

- Preprocessing methods for multiple spectra *de novo* peptide sequencing.

  There are already a lot of preprocessing methods developed for traditionally used CID spectra, but not many for multiple spectra sequencing techniques. One key preprocessing step in these techniques is spectra merging. New methods can focus on the improvement of merging criteria for the pairs of spectra studied in this thesis, or the merging of other kinds of spectra combinations. In addition, effective denoising methods for these spectra can be another potential research topic.

- Software development for the proposed *de novo* peptide sequencing methods.

  A series of computational methods are presented in this thesis, and it would be helpful to develop integrated software packages for interested users. All presented methods can be integrated into one single software package to be applicable for peptide sequencing of different types of spectra or spectra combinations.

References

[1] Bin Ma and Richard Johnson. *De novo* sequencing and homology searching. *Molecular & Cellular Proteomics*, 11(2):1–16, 2012.

[2] Ingvar Eidhammer, Kristian Flikka, Lennart Martens, and Svein-Ole Mikalsen. *Computational methods for mass spectrometry proteomics*. John Wiley & Sons, 2008.

[3] Michael Kinter and Nicholas E Sherman. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. John Wiley & Sons, 2000.

[4] Katheryn A Resing and Natalie G Ahn. Proteomics strategies for protein identification. *FEBS Letters*, 579:885–889, 2005.

[5] Richard S Johnson and Klaus Biemann. Computer program (DEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomedical & Environmental Mass Spectrometry*, 18(11):945–957, 1989.

[6] Leo McHugh and Jonathan W Arthur. Computational methods for protein identification from mass spectrometry data. *PLoS Computational Biology*, 4(2), 2008.

[7] Ruedi Aebersold and David R Goodlett. Mass spectrometry in proteomics. *Chemical Reviews*, 101(2):269–296, 2001.

[8] Vicki H Wysocki, Katheryn A Resing, Qingfen Zhang, and Guilong Cheng. Mass spectrometry of peptides and proteins. *Methods*, 35(3):211–222, 2005.

[9] Fred W McLafferty. *Interpretation of Mass Spectra*. University Science Books, 1993.

[10] James J Pitt. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical Biochemist Reviews*, 30(1):19, 2009.

[11] Alan G Marshall, Christopher L Hendrickson, and George S Jackson. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrometry Reviews*, 17(1):1–35, 1998.

[12] Raymond E March. Quadrupole ion trap mass spectrometry: theory, simulation, recent developments and applications. *Rapid Communications in Mass Spectrometry*, 12(20):1543–1554, 1998.

[13] Seungjin Na, Eunok Paek, and Cheolju Lee. Structural characterization of peptides via tandem mass spectrometry of their dilithiated monocations. *Analytical Chemistry*, 80(5):1520–1528, 2008.

[14] Changjiang Xu and Bin Ma. Software for computational peptide identification from MS/MS data. *Drug Discovery Today*, 11(13):595–600, 2006.

[15] Bobbie-Jo M Webb-Robertson and William R Cannon. Current trends in computational inference from mass spectrometry-based proteomics. *Briefings in Bioinformatics*, 8(5):304–317, 2007.

[16] Michael R Hoopmann and Robert L Moritz. Current algorithmic solutions for peptide-based proteomics data generation and identification. *Current Opinion in Biotechnology*, 24(1):31–38, 2013.

[17] Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. NovoGMET:*De novo* peptide sequencing using graphs with multiple edge types (GMET) for ETD/ECD spectra. In *The 10th International Symposium on Bioinformatics Research and Applications (ISBRA2014)*, pages 200–211, 2014.

[18] Hao Chi, Kun He, Bing Yang, Zhen Chen, Rui-Xiang Sun, Sheng-Bo Fan, Kun Zhang, Chao Liu, Zuo-Fei Yuan, Quan-Hui Wang, Si-Qi Liu, Meng-Qiu Dong, and Si-Min He. pFind-Alioth: a novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data. *Journal of Proteomics*, 125(0):89 – 97, 2015.

[19] Yaojun Wang, Fei Yang, Peng Wu, Dongbo Bu, and Shiwei Sun. OpenMS-Simulator: an open-source software for theoretical tandem mass spectrum prediction. *BMC Bioinformatics*, 16(1):110:1–6, 2015.

[20] Jimmy K Eng, Brian C Searle, Karl R Clauser, and David L Tabb. A face in the crowd: recognizing peptides through database search. *Molecular & Cellular Proteomics*, 10(11):R111–009522, 2011.

[21] Chongle Pan, Byung Park, William McDonald, Patricia Carey, Jillian Banfield, Nathan VerBerkmoes, Robert Hettich, and Nagiza Samatova. A high-throughput *de novo* sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC Bioinformatics*, 11(1):118, 2010.

[22] David L Tabb, Ze-Qiang Ma, Daniel B Martin, Amy-Joan L Ham, and Matthew C Chambers. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *Journal of Proteome Research*, 7(9):3838–3846, 2008.

[23] Marshall Bern and David Goldberg. *De novo* analysis of peptide tandem mass spectra by spectral graph partitioning. *Journal of Computational Biology*, 13(2):364–378, 2006.

[24] Peter A DiMaggio and Christodoulos A Floudas. *De novo* peptide identification via mixed-integer linear optimization and tandem mass spectrometry. *Computer Aided Chemical Engineering*, 24:989–994, 2007.

[25] Bingwen Lu and Ting Chen. Algorithms for *de novo* peptide sequencing using tandem mass spectrometry. *Drug Discovery Today: Biosilico*, 2(2):85–90, 2004.

[26] Vlado Dancik, Theresa A Addona, Karl R Clauser, James E Vath, and Pavel A Pevzner. *De novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.

[27] Rainer Cramer. Editorial for "advances in biological mass spectrometry and proteomics". *Methods*, 54:349–350, 2011.

[28] Ingvar Eidhammer, Kristian Flikka, Lennart Martens, and Svein-Ole Mikalsen. *Computational methods for mass spectrometry proteomics*. John Wiley & Sons, 2008.

[29] Vicki H Wysocki, George Tsaprailis, Lori L Smith, and Linda A Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, 35(12):1399–1406, 2000.

[30] Alex G Harrison. To b or not to b: the ongoing saga of peptide b ions. *Mass Spectrometry Reviews*, 28(4):640–654, 2009.

[31] Lin He and Bin Ma. ADEPTS: advance peptide *de novo* sequencing with a pair of tandem mass spectra. *Journal of Bioinformatics and Computational Biology*, 8:981–994, 2010.

[32] Maria Fälth, Mikhail M Savitski, Michael L Nielsen, Frank Kjeldsen, Per E Andren, and Roman A Zubarev. Analytical utility of small neutral losses from reduced species in electron capture dissociation studied using SwedECD database. *Analytical Chemistry*, 80(21):8089–8094, 2008.

[33] Robert J Chalkley, Katalin F Medzihradszky, Aenoch J Lynn, Peter R Baker, and Alma L Burlingame. Statistical analysis of peptide electron transfer dissociation fragmentation mass spectrometry. *Analytical Chemistry*, 82(2):579–584, 2010.

[34] Alan L Gray, John G Williams, Ahmet T Ince, and Martin Liezers. Communication. noise sources in inductively coupled plasma mass spectrometry: an investigation of their importance to the precision of isotope ratio measurements. *Journal of Analytical Atomic Spectrometry*, 9:1179–1181, 1994.

[35] Adrian Guthals and Nuno Bandeira. Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Molecular & Cellular Proteomics*, 11(9):550–557, 2012.

[36] Annette Michalski, Nadin Neuhauser, Jürgen Cox, and Matthias Mann. A systematic investigation into the nature of tryptic HCD spectra. *Journal of Proteome Research*, 11(11):5479–5491, 2012.

[37] John EP Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9528–9533, 2004.

[38] Roman A Zubarev, Neil L Kelleher, and Fred W McLafferty. Electron capture dissociation of multiply charged protein cations. A nonergodic proces. *Journal of the American Chemical Society*, 120(16):3265–3266, 1998.

[39] Leann M Mikesh, Beatrix Ueberheide, An Chi, Joshua J Coon, John EP Syka, Jeffrey Shabanowitz, and Donald F Hunt. The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1764(12):1811 – 1822, 2006.

[40] Julia Wiesner, Thomas Premsler, and Albert Sickmann. Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics*, 8(21):4466–4483, 2008.

[41] Min-Sik Kim and Akhilesh Pandey. Electron transferm dissociation mass spectrometry in proteomics. *Proteomics*, 12(4-5):530–42, 2012.

[42] Rovshan G Sadygov, Daniel Cociorva, and John R Yates. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature Methods*, 1(3):195–202, 2004.

[43] Jimmy K Eng, Ashley L McCormack, and John R Yates. 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.

[44] Matthias Mann and Matthias Wilm. Error-tolerant identification of peptides in sequence tags. *Analytical Chemistry*, 66(24):4390–4399, 1994.

[45] Bassil I Dahiyat and Stephen L Mayo. *De novo* protein design: fully automated sequence selection. *Science*, 278:82–87, 1997.

[46] Fang-Xiang Wu, Pierre Gagne, Arnaud Droit, and Guy G Poirier. RT-PSM, a real-time program for peptide-spectrum matching with statistical significance. *Rapid Communications in Mass Spectrometry*, 20(8):1199–1208, 2006.

[47] Mikhail M Savitski, Michael L Nielsen, and Roman A Zubarev. New data base-independent, sequence tag-based scoring of peptide MS/MS data validates mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Molecular & Cellular Proteomics*, 4(8):1180–1188, 2005.

[48] Bingwen Lu and Ting Chen. Algorithms for *de novo* peptide sequencing using tandem mass spectrometry. *BioSilico*, 2:85–90, 2004.

[49] Neil C Jones and Pavel A Pevzner. *An Introduction to Bioinformatics Algorithms*. Cambridge, Massachusetts: MIT press, 2004.

[50] Tsuneaki Sakurai, Kenji Matsuo, Hideo Matsuda, and Ituso Katakuse. Paas 3: a computer program to determine probable sequence of peptides from mass spectrometric data. *Biological Mass Spectrometry*, 11(8):396–399, 1984.

[51] Hubert A Scoble, James E Biller, and Klaus Biemann. A graphics display-oriented strategy for the amino acid sequencing of peptides by tandem mass spectrometry. *Fresenius Journal of Analytical Chemistry*, 327(2):239–245, 1987.

[52] Yan Yan, Shenggui Zhang, and Fang-Xiang Wu. Applications of graph theory in protein structure identification. *Proteome Science*, 9(Suppl 1):S17, 2011.

[53] Christian Bartels. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomedical & Environmental Mass Spectrometry*, 19:363–368, 1990.

[54] Wade M Hines, Arnold M Falick, Alma L Burlingame, and Bradford W Gibson. Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *Journal of the American Society for Mass Spectrometry*, 3(4):326–336, 1992.

[55] Jorge Fernández-de Cossio, Javier Gonzalez, and Vladimir Besada. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Computer Applications in the Biosciences: CABIOS*, 11(4):427–434, 1995.

[56] Vlado Dančík, Theresa A Addona, Karl R Clauser, and James E Vath. *De novo* peptide sequencing via tandem mass spectrometry: a graph-theoretical approach. In *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, pages 135–144. ACM, 1999.

[57] Ting Chen, Mingyang Kao, Matthew Tepel, John Rush, and George M. A dynamic programming approach for *de novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 8:325–337, 2001.

[58] Lijuan Mo, Debojyoti Dutta, Yunhu Wan, and Ting Chen. MsNovo: a dynamic programming algorithm for *de novo* peptide sequencing via tandem mass spectrometry. *Analytical Chemistry*, 79(13):4870–4878, 2007.

[59] Ari Frank and Pavel A Pevzner. PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77:964–973, 2005.

[60] Sergey Pevtsov, Irina Fedulova, Hamid Mirzaei, Charles Buck, and Xiang Zhang. Performance evaluation of existing *de novo* sequencing algorithms. *Journal of Proteome Research*, 5(11):3018–3028, 2006.

[61] Bernd Fischer, Volker Roth, Franz Roos, Jonas Grossmann, Sacha Baginsky, Peter Widmayer, Wilhelm Gruissem, and Joachim M Buhmann. NovoHMM: a hidden markov model for *de novo* peptide sequencing. *Analytical Chemistry*, 77(22):7265–7273, 2005.

[62] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–2342, 2003.

[63] Bin Ma, Kaizhong Zhang, and Chengzhi Liang. An effective algorithm for peptide *de novo* sequencing from MS/MS spectra. *Journal of Computer and System Sciences*, 70(3):418–430, 2005.

[64] Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A Lajoie, and Bin Ma. PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics*, 11(4):M111–010587, 2012.

[65] Jingfen Zhang, Simin He, Charles X Ling, Xingjun Cao, Rong Zeng, and Wen Gao. PeakSelect: preprocessing tandem mass spectra for better peptide identification. *Rapid Communications in Mass Spectrometry*, 22(8):1203–1212, 2008.

[66] Viswanadham Sridhara, Dina L Bai, An Chi, Jeffrey Shabanowitz, Donald F Hunt, Stephen H Bryant, and Lewis Y Geer. Increasing peptide identifications and decreasing search times for ETD spectra by pre-processing and calculation of parent precursor charge. *Proteome Science*, 10(1):1–10, 2012.

[67] Bo Yan, Chongle Pan, Victor N Olman, Robert L Hettich, and Ying Xu. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. *Bioinformatics*, 21(5):563–574, 2005.

[68] Bin Ma. Challenges in computational analysis of mass spectrometry data for proteomics. *Journal of Computer Science and Technology*, 25(1):107–123, 2010.

[69] Chongle Pan, Byung H Park, William H McDonald, Patricia A Carey, Jillian F Banfield, Nathan C VerBerkmoes, Robert L Hettich, and Nagiza F Samatova. A high-throughput *de novo* sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC Bioinformatics*, 11(1):118, 2010.

[70] Sara Brunetti, Elena Lodi, Elisa Mori, and Maria Stella. PARPST: a parallel algorithm to find peptide sequence tags. *BMC Bioinformatics*, 9(Suppl 4):S11, 2008.

[71] Arnie M Falick, William M Hines, Katalin F Medzihradszky, Matthew A Baldwin, and Bradford W Gibson. Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 4(11):882 – 893, 1993.

[72] Hao Chi, Rui-Xiang Sun, Bing Yang, Chun-Qing Song, Le-Heng Wang, Chao Liu, Yan Fu, Zuo-Fei Yuan, Hai-Peng Wang, Si-Min He, and Meng-Qiu Dong. pNovo: *de novo* peptide sequencing and identification using HCD spectra. *Journal of Proteome Research*, 9(5):2713–2724, 2010.

[73] Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. NovoHCD: *De novo* peptide sequencing from HCD spectra. *IEEE Transactions on NanoBioscience*, 13(2):65–72, June 2014.

[74] Yufeng Shen, Nikola Tolić, Fang Xie, Rui Zhao, Samuel O Purvine, Athena A Schepmoes, J. Moore, Ronald, Gordon A Anderson, and Richard D Smith. Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. *Journal of Proteome Research*, 10(9):3929–3943, 2011.

[75] Hao Chi, Haifeng Chen, Kun He, Long Wu, Bing Yang, Rui-Xiang Sun, Jianyun Liu, Wen-Feng Zeng, Chun-Qing Song, Si-Min He, and Meng-Qiu Dong. pNovo+: *De novo* peptide sequencing using complementary HCD and ETD tandem mass spectra. *Journal of Proteome Research*, 12(2):615–625, 2013.

[76] Christian K Frese, AF Maarten Altelaar, Marco L Hennrich, Dirk Nolting, Martin Zeller, Jens Griep-Raming, Albert JR Heck, and Shabaz Mohammed. Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-orbitrap velos. *Journal of Proteome Research*, 10(5):2377–2388, 2011.

[77] Andreas Bertsch, Andreas Leinenbach, Anton Pervukhin, Markus Lubeck, Ralf Hartmer, Carsten Baessmann, Yasser Abbas Elnakady, Rolf Müller, Sebastian Böcker, Christian G Huber, and Oliver Kohlbacher. *De novo* peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis*, 30:3736–3747, 2009.

[78] Roman A Zubarev. Reactions of polypeptide ions with electrons in the gas phase. *Mass Spectrometry Reviews*, 22(1):57–77, 2003.

[79] Mikhail M Savitski, Michael L Nielsen, Frank Kjeldsen, and Roman A Zubarev. Proteomics-grade *de novo* sequencing approach. *Journal of Proteome Research*, 4(6):2348–2354, 2005.

[80] Ritendra Datta and Marshall Bern. Spectrum fusion: using multiple mass spectra for *de novo* peptide sequencing. *Journal of Computational Biology*, 16(8):1169–1182, 2009.

[81] Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. NovoPair: *De novo* peptide sequencing for tandem mass spectra pair. In *The IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 150–155, 2014.

[82] Maarten A F Altelaar, Danny Navarro, Jos Boekhorst, Bas van Breukelen, Berend Snel, Shabaz Mohammed, and Albert J R Heck. Database independent proteomics analysis of the ostrich and human proteome. *Proceedings of the National Academy of Sciences*, 109(2):407–412, 2012.

[83] Adrian Guthals, Karl R Clauser, Ari M Frank, and Nuno Bandeira. Sequencing-grade *de novo* analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides. *Journal of Proteome Research*, 12(6):2846–2857, 2013.

[84] Kyowon Jeong, Sangtae Kim, and Pavel A Pevzner. UniNovo: a universal tool for *de novo* peptide sequencing. *Bioinformatics*, 29(16):1953–1962, 2013.

[85] Marshall Bern, Yuhan Cai, and David Goldberg. Lookup peaks: a hybrid of *de novo* sequencing and database search for protein identification by tandem mass spectrometry. *Analytical Chemistry*, 79(4):1393–1400, 2007.

[86] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert JR Heck, and Pavel A Pevzner. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics*, 9(12):2840–2852, 2010.

[87] Katalin F. Medzihradszky and Robert J. Chalkley. Lessons in *de novo* peptide sequencing by tandem mass spectrometry. *Mass Spectrometry Reviews*, 34(1):43–63, 2015.

[88] Jiarui Ding, Jinhong Shi, Guy Poirier, and Fang-Xiang Wu. A novel approach to denoising ion trap tandem mass spectra. *Proteome Science*, 7(1):9, 2009.

[89] Wenjun Lin, Fang-Xiang Wu, Jinhong Shi, Jiarui Ding, and Wenjun Zhang. An adaptive approach to denoising tandem mass spectra. *Proteomics*, 11(19):3773–3778, 2011.

[90] Fang-Xiang Wu, Pierre Gagné, Arnaud Droit, and Guy G Poirier. Quality assessment of peptide tandem mass spectra. *BMC Bioinformatics*, 9(Suppl 6):S13, 2008.

[91] Seungjin Na, Eunok Paek, and Cheolju Lee. CIFTER: automated charge-state determination for peptide tandem mass spectra. *Analytical Chemistry*, 80:1520–1528, 2008.

[92] Jinhong Shi and Fang-Xiang Wu. Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation. *Proteome Science*, 9(Suppl 1):S3, 2011.

[93] An-Min Zou, Jinhong Shi, Jiarui Ding, and Fang-Xiang Wu. Charge state determination of peptide tandem mass spectra using support vector machine (SVM). *IEEE Transactions on Information Technology in Biomedicine*, 14(3):552–558, 2010.

[94] Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. A multi-edge graph based *de novo* peptide sequencing method for HCD spectra. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 176–181, December 2013.

[95] Alexey V Nefedov, Indranil Mitra, Allan R Brasier, and Rovshan G Sadygov. Examining troughs in the mass distribution of all theoretically possible tryptic peptides. *Journal of Proteome Research*, 10(9):4150–4157, 2011.

[96] Thomas A Hansen, Fedor Kryuchkov, and Frank Kjeldsen. Reduction in database search space by utilization of amino acid composition information from electron transfer dissociation and higher-energy collisional dissociation mass spectra. *Analytical Chemistry*, 84(15):6638–6645, 2012.

[97] Anton A Goloborodko, Mikhail V Gorshkov, David M Good, and Roman A Zubarev. Sequence scrambling in shotgun proteomics is negligible. *Journal of the American Society for Mass Spectrometry*, 22(7):1121–1124, 2011.

[98] Christian K Frese, Maarten A F Altelaar, Marco L Hennrich, Dirk Nolting, Martin Zeller, Jens Griep-Raming, Albert J R Heck, and Shabaz Mohammed. Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-orbitrap velos. *Journal of Proteome Research*, 10(5):2377–2388, 2011.

[99] Matrix science. `http://www.matrixscience.com/`, 2013.

[100] Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. NovoExD: *De novo* peptide sequencing for ETD/ECD spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press.

[101] Jason M Hogan, Roger Higdon, Natali Kolker, and Eugene Kolker. Charge state estimation for tandem mass spectrometry proteomics. *Omics: A Journal of Integrative Biology*, 9(3):233–250, 2005.

[102] Aaron A Klammer, Christine C Wu, Michael J MacCoss, and William Stafford Noble. Peptide charge state determination for low-resolution tandem mass spectra. In *Computational Systems Bioinformatics Conference, 2005*, pages 175–185. IEEE, 2005.

[103] Shiwei Sun, Chungong Yu, Yantao Qiao, et al. Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. *Journal of Proteome Research*, 7(1):202–208, 2008.

[104] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[105] Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. A framework of *de novo* peptide sequencing for multiple tandem mass spectra. *IEEE Transactions on NanoBioscience*, 14(4):478–484, June 2015.

[106] Lin He, Xi Han, and Bin Ma. *De novo* sequencing with limited number of post-translational modifications per peptide. *Journal of Bioinformatics and Computational Biology*, 11(04):1350007: 1–12, 2013.

Appendix A

List of Publications

The followings are the thesis-related journal publications.

1. Yan Yan, Shenggui Zhang, and Fang-Xiang Wu. Applications of graph theory in protein structure identification. *Proteome Science*, 9(Suppl 1):S17, 2011.

2. Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. NovoHCD: *De novo* peptide sequencing from HCD spectra. *IEEE Transactions on NanoBioscience*, 13(2):65–72, June 2014.

3. Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. NovoExD: *De novo* peptide sequencing for ETD/ECD spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press.

4. Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. A framework of *de novo* peptide sequencing for multiple tandem mass spectra. *IEEE Transactions on NanoBioscience*, 14(4):478–484, June 2015.

5. Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. Recent developments in computational methods for *de novo* peptide sequencing via tandem mass spectrometry (MS/MS). *Protein & Peptide Letters*, August, 2015 (Accepted).

6. Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. *De novo* peptide sequencing using CID and HCD spectra pairs. Submitted to *Proteomics*, 2015.

The followings are the thesis-related conference publications.

1. Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. A multi-edge graph based *de novo* peptide sequencing method for HCD spectra. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 176–181, December 2013.

2. Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. NovoGMET:*De novo* peptide sequencing using graphs with multiple edge types (GMET) for ETD/ECD spectra. In *The 10th International Symposium on Bioinformatics Research and Applications (ISBRA2014)*, pages 200–211, 2014.

3. Yan Yan, Anthony J Kusalik, and Fang-Xiang Wu. NovoPair: *De novo* peptide sequencing for tandem mass spectra pair. In *The IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 150–155, 2014.

The following is the thesis-unrelated publication.

1. Bolin Chen, Yan Yan, Jinhong Shi, Shenggui Zhang, and Fang-Xiang Wu. An improved graph entropy-based method for identifying protein complexes. In *2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 123–126. IEEE, 2011.

Appendix B

Copyright Permissions

Copyright forms of thesis-related publications are attached in the following pages.

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.