Effective Strategies for Improving Peptide Identification with Tandem Mass Spectrometry

by

Xi Han

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

Waterloo, Ontario, Canada, 2011

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

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

Abstract

Tandem mass spectrometry (MS/MS) has been routinely used to identify peptides from protein mixtures in the field of proteomics. However, only about 30% to 40% of current MS/MS spectra can be identified, while many of them remain unassigned, even though they are of reasonable quality.

The ubiquitous presence of post-translational modifications (PTMs) is one of the reasons for current low spectral identification rate. In order to identify post-translationally modified peptides, most existing software requires the specification of a few possible modifications. However, such knowledge of possible modifications is not always available. In this thesis, we describe a new algorithm for identifying modified peptides without requiring users to specify the possible modifications before the search routine; instead, all modifications from the Unimod database are considered. Meanwhile, several new techniques are employed to avoid the exponential growth of the search space, as well as to control the false discoveries due to this unrestricted search approach. A software tool, PeaksPTM, has been developed and it has already achieved a stronger performance than competitive tools for unrestricted identification of post-translationally modified peptides.

Another important reason for the failure of the search tools is the inaccurate mass or charge state measurement of the precursor peptide ion. In this thesis, we study the precursor mono-isotopic mass and charge determination problem, and propose an algorithm to correct precursor ion mass error by assessing the isotopic features in its parent MS spectrum. The algorithm has been tested on two annotated data sets and achieved almost 100 percent accuracy. Furthermore, we have studied a more complicated problem, the MS/MS preprocessing problem, and propose a spectrum deconvolution algorithm. Experiments were provided to compare its performance with other existing software.

Acknowledgements

I am very grateful to my supervisor, Professor Bin Ma, for his guidance, outstanding support and invaluable advice on my research topics. His knowledge, expertise and deep insights in this research field greatly improve my research skills and prepare me for future challenges.

I would like to express my appreciation to my collaborators: Baozhen Shan and Lei Xin in Bioinformatics Solutions Inc., for many valuable discussions and suggestions, and their kindly helps to find experimental data.

Special thanks to my committee members, Professor Ming Li and Professor Forbes Burkowski, for taking their precious time to review my thesis and provide valuable feedback and suggestions.

My thanks would go to my parents, my sister and my husband, who are always there for me. Their support, understanding and endless love are the source of my courage and strength to conquer any difficulty in my study and life.

Last but not least, I would like to thank my dear friends, Ying Liu, Shuaicheng Li, Xuefeng Cui, Wei He, Jiewen Wu, Tiantian Bian, Laleh Soltan Ghoraie, as well as many friends who enrich my life and have made my stay at Waterloo really enjoyable.

Dedication

This is dedicated to my parents, my sister, and my husband.

Table of Contents

List of Tables **ix**

List of Figures **xii**

1 Introduction **1**

1.1 Motivation 1

1.2 Research Objectives 2

1.3 Overview of the Thesis 3

2 Background **4**

2.1 Mass Spectrometry (MS) Based Proteomics 4

2.1.1 Proteomics 4

2.1.2 Mass Spectrometry Technology 5

2.1.3 MS/MS-Based Shotgun Proteomics Strategy 7

2.2 Peptide Sequence Assignment to MS/MS Spectra 9

2.2.1 Database Search 9

2.2.2 *De Novo* Sequencing 10

2.3 Unrestricted PTM Search 10

2.3.1 Preliminary Knowledge About PTM and Modified Positions 11

2.3.2 Error Rate Estimation for Peptide Spectrum Matches 13

2.3.3 Target-Decoy Strategy for False Discovery Rate Estimation 13

2.4 Two Related Problems for Peptide Identification 15
2.4.1 Preliminary Knowledge for Mass Spectrometry Data Analysis . . . 15
2.4.2 MS/MS Spectra Preprocessing 16

3 Unrestrictive PTM search **18**
3.1 Method 18
3.1.1 Method Overview 19
3.1.2 Protein Identification 20
3.1.3 Single-PTM Peptide Candidate Search 21
3.1.4 Peptide Rescoring 22
3.1.5 Estimation of the False Discovery Rate 27
3.2 Experiments and Results 28
3.2.1 Data Sets 29
3.2.2 Cross-Training on Two Data Sets 29
3.2.3 Comparison Between Multiple Search Engines 30
3.2.4 Comparison with MOD^i 32
3.2.5 Consensus Strategy and Analysis 34
3.2.6 Summary of Identified PTMs 36
3.3 Discussion 36

4 Two Related Problems for Peptide Identification **40**
4.1 Precursor Mono-Isotopic Mass and Charge Correction Problem 40
4.1.1 Method Overview 41
4.1.2 Candidate Generation 42
4.1.3 Candidate Evaluation 47
4.2 MS/MS Spectrum Deconvolution Problem 48
4.2.1 MS/MS Spectra Deconvolution Problem: An Application of Charge and Mass Correction Algorithm 48

4.2.2 Method Overview 50
4.2.3 Isotopic Envelopes Selection 52
4.2.4 Fitting Score for Choosing the Best Candidates 52
4.2.5 Multiple Charge Deconvolution 54
4.3 Experiments and Results 55
4.3.1 Experiments for Precursor Mono-Isotopic Mass and Charge Correction Algorithm 55
4.3.2 Results 55
4.3.3 Experiments for MS/MS Spectra Deconvolution Algorithm Compared with Two Other Software Tools 57
4.3.4 Results 58
4.4 Discussion 60

5 Conclusion and Future Work **63**
5.1 Conclusion 63
5.2 Future Work 64

References **69**

List of Tables

2.1 Mass resolution and accuracy of mass analyzers. 6

3.1 An example of peptide pairs in the identification results. Peptides identified for MS/MS scan No.8111 and No.11626 compose a peptide pair. 23

3.2 The summary of 29 PTMs which are frequently reported by previous experiments. 25

3.3 The numbers of identified peptides with 1% FDR under different settings of training and testing data sets. 29

3.4 The numbers of unique modified peptides containing the most common PTMs in the human heart data set. 37

4.1 The look-up table for the prediction of isotopic abundance distribution. The mono-isotopic masses in this table range from 50 Da to 75,000 Da. 44

4.2 The comparisons of correctly and partial correctly identified peptides by using a *de novo* sequencing approach on three sets of the pre-processed spectra. 58

List of Figures

2.1 The basic components of a mass spectrometer: an ion source, a mass analyzer and a mass detector. The relative abundances of measured ions are reported in mass spectra. 5

2.2 Fragmentation sites of a peptide and fragment ions produced. a-, b-, and c-type ions contain the N-terminus; x-, y-, and z- ions contain the C-terminus; R_1, R_2, R_3 and R_4 represent the side chains of amino acid residues. 7

2.3 A typical experiment procedure for protein complex identification using tandem mass spectrometry [1]. 8

2.4 Database search strategy for peptide identification. 9

2.5 *De novo* sequencing strategy for peptide identification. 10

2.6 Three kinds of positions at which a PTM may occur on a protein or peptide. Circle attached on NH_2- group represents a PTM added on N-terminus. Circle on -$COOH$ group represents a PTM added on C-terminus. Triangle attached on R_3 stands for a PTM added on side chain of a residue, and R_1, R_2, R_3, and R_4 are all possible modified positions, depending on the PTM's target residue type. 12

2.7 An example of an isotopic pattern. Peak with m/z 323.15 Da is the monoisotopic peak of this isotopic envelope. Peaks with m/z 324.15 Da and 325.15 Da are two isotopes. The three peaks together compose an isotopic pattern of an unknown ion. 16

2.8 An example of different charged isotopic envelopes of the same fragmented ion. Peak list {162.08, 162.58, 163.08} is its isotope pattern with charge 2; peak list {323.15, 324.15, 325.15} is its isotope pattern with charge 1. The inner distance $d = 1/z$, where z is the charge state that equals 1 or 2. . . 17

3.1 The LDF score distributions of single-PTM peptides identified from target sequences and decoy sequences, respectively. (a) The distribution with peptide pairs, and (b) without peptide pairs. Modified peptides from the target sequences tend to have more peptide pairs than those from the decoy sequences. 23

3.2 The LDF score distributions of the peptide candidates identified with no PTM, a common PTM, and a rare PTM, from (a) the target and (b) decoy proteins. 26

3.3 The modified target-decoy strategy used in PeaksPTM. P_1 is the set of proteins coming from the target database, while P_2 is from the decoy database. 27

3.4 An example of the peptide level shuffled approach used in PeaksPTM. . . . 28

3.5 The comparison of reported modified peptides by InsPecT, Mascot, Paragon and PeaksPTM. The curves show the relation between the estimated FDR and the number of results reported. 31

3.6 A large portion of PeaksPTM's high confidence (FDR $\leq 1\%$) modified PSMs are also identified by at least one other engine, either with high or low confidence. 32

3.7 The comparison of PeaksPTM, MOD^i and InsPecT on the reduced database with 10 target + 10 decoy proteins. The curves show the relation between the estimated FDR and the number of results reported. 33

3.8 The consensus result of identified PSMs by using consensus strategy on human heart data set. 34

3.9 The Venn diagram shows the composition of confidently identified modified PSMs by the four search engines using a consensus strategy. 35

4.1 An example of precursor mass shift. 41

4.2 Peak centroiding. 42

4.3 Examples of noise peaks and isotopic peaks. (a), (b) and (c) show spurious peak groups, and (d) gives an example of real isotopic envelope. 43

4.4 An example: steps to determine the mono-isotopic mass and charge state. . 49

4.5 The frame work of the proposed deconvolution algorithm. 51

4.6 Experiment on data set A: precursor mono-isotopic peak correction compared with human expert's annotation: (a) before correction vs. (b) after correction; experiment on data set B: (c) before correction vs. (d) after correction. 56

4.7 The number of matched ions found by increasing the number of selected peaks in each spectrum. 59

4.8 An example to show that our algorithm can handle the overlapping problem. (a) The observed peak distributions within an m/z region of a tandem MS spectrum. (b) Peaks intensity distribution after centroiding. From this distribution, our algorithm identified two overlapped isotopic envelopes (E_1 and E_2) with charge state 2. E_1 started from the peak with m/z 514.75 Da (c), and E_2 started from the highest peak with m/z 515.25 Da (d). (c) and (d) also illustrates their assigned intensities, respectively. (e) The isotope distribution generated by the overlapping E_1 and E_2. (f) A comparison to show that the identified peak distribution by our algorithm is very close to the observed peak distribution. E_1 and E_2 were found matching theoretical ions: $\{y_9 - NH_3\}$-ion and $\{y_9 - H_2O\}$-ion, respectively. 61

Chapter 1

Introduction

1.1 Motivation

In the past twenty years, tandem mass spectrometry (MS/MS) has become the method of choice for peptide identification in the field of proteomics. In a typical bottom-up proteomics analysis, the enzymatically digested peptides from a protein mixture are measured with a LC-MS/MS experiment to produce a large number of MS/MS spectra. Each spectrum is compared with the peptides in a protein sequence database to find the best match. Many software tools have been developed for peptide identifications from MS/MS data. The most common tools include PEAKS [2], Mascot [3], Sequest [4], X!Tandem [5], and OMSSA [6]. However, in most cases, the portion of identified peptides from the current MS/MS spectra is quite low, around 30% to 40%. Many MS/MS spectra remain unassigned, even though they are of reasonable quality.

There are several reasons to explain the low identification rate: constrained database search parameters (e.g. searching for tryptic peptides only), inaccurate charge state or mass measurement of the precursor peptide ion, the presence of chemical or post-translational modifications (PTMs) not considered in the search, and incompleteness of the searched protein sequence database [7] [8] [9] [10].

The limited support for peptides with PTMs in the current database search tools is believed to be one of the major reasons for the low characterization rate of the MS/MS spectra [11]. The peptide identification approach for a conventional database search was first proposed by Yates et al. [12] and generally achieved by the following procedure: A human user first specifies the PTM types expected to be seen in the results. If a PTM is

specified as a fixed modification (such as carbamidomethylation on cysteine), every occurrence of its target residue will be replaced with a modified residue. A fixed modification will not affect the software's running time. However, if a PTM is specified as a variable one(such as phosphorylation on serine, threonine, or tyrosine), each applicable residue in the sequence database will be tried in two different ways (with or without the modification), which increases the running time. In particular, specifying several variable PTMs creates multiple possible modification sites for an average peptide, causing an exponential growth of search space. This growth will increase not only the running time to an unacceptable level, but also the potential for false discoveries. As a result, when a conventional search engine is used for peptide identification, only a few variable PTMs can be practically specified. Those unspecified PTMs are lost because of the limitations of the software. Today, hundreds of PTMs have been found and characterized. The Unimod PTM database [13] lists more than 500 entries and the DeltaMass database [14] includes over 300. Recent research work [15] shows that most eukaryotic proteins are post-translationally modified. The identification of modified proteins, as well as the PTM types and modification sites on the proteins, is essential to a thorough understanding of the biological functions of PTMs and is of great interest for proteomics research [4] [5] [16] [17].

Another important reason for the failure of the database search tools to identify corresponding peptides is the inaccurate charge state or mass measurement of the precursor peptide ion. Due to the incapability of recognizing the mono-isotope peak of the precursor ion, even high-resolution tandem mass spectrometer such as LTQ-Orbitrap often reports the precursor mass one or more Daltons (Da) from the correct value. The resolution of these high resolution instrument can achieve 10 ppm (parts per million). Suppose there is a peptide with mass value 2000 Da, the error tolerance of 10 ppm of 2000 Da is 0.02 Da, which is 50 fold smaller than one isotope offset (1 Da). The default setting of error tolerance of precursor mass would cause the software analysis to fail unless, contrary to the nature of high-resolution experiments, a bigger mass error tolerance, at least 500 ppm, is used. However, this would lead to a sharp increase on the number of the theoretical peptides needed to be compared with each spectrum, consequently increasing the running time of the search program.

1.2 Research Objectives

The objective of this thesis is to study the strategies that improve the typical data analysis workflow to increase both the quantity and the accuracy of identified peptides from MS/MS data. The possible strategies are studied with respect to two goals:

- Design an unrestricted PTM search tool to increase the number of identified peptides containing PTMs. On one hand, the unrestricted PTM search strategy should not require any prior knowledge of possible PTMs in the data set. Therefore, it can overcome the drawback of conventional database search which is blind to modified peptides if the PTMs are not specified by the user before the search. On the other hand, it should prevent the combinatorial explosion of the search space and the introduction of false discoveries, both of which are caused by the increasing number of possible PTMs.

- Design algorithms to refine the spectra data before the routine database search and *de novo* sequencing. The first algorithm is designed to correct the mass and charge value of the precursor ion. The mono-isotopic mass of the precursor ion is crucial for most existing software to identify a peptide from its MS/MS spectrum. The correction algorithm will enable the usage of a small precursor error tolerance in both database search and *de novo* sequencing. Based on the first algorithm, the second algorithm is designed to preprocess the MS/MS data, in term of enhancing the real ion signals relative to the noise signals.

1.3 Overview of the Thesis

This thesis is organized as follows. Fundamentals for MS-based proteomics, such as database search approach, unrestricted PTM search, false discovery rate assessment, as well as preliminary knowledge about MS data analysis and preprocessing are introduced in Chapter 2. Chapter 3 presents the details of the design of PeaksPTM, an MS-based unrestricted PTM search tool. Experiments are provided to compare its performance with four other unrestricted PTM search tools. We also adopt a consensus strategy to combine all the results of each tool, and discuss the important impact of modified peptides towards the increase of identification rate of MS/MS data. This chapter is based on our previous publication [18]. Chapter 4 defines the precursor mono-isotopic mass and charge determination problem and the MS/MS spectra preprocess problem. Two efficient algorithms are proposed, and their performances are studied. Finally, Chapter 5 gives a conclusion and proposes future work.

Chapter 2

Background

2.1 Mass Spectrometry (MS) Based Proteomics

2.1.1 Proteomics

Proteomics is the large-scale study of proteins, particularly their structures and functions [19] [20] [21]. The word "proteome" comes from a blend of "protein" and "genome", and was first used in 1995 [22] to describe the protein complement of a genome. In the past twenty years, the proteome was imperceptibly transmuted into a new discipline: proteomics. It focuses on the large scale study of protein properties, such as expression level, post-translational modification, protein-protein interactions etc., and gives a global view of biological processes, disease and networks at the protein level [19].

In the study of proteomics, a number of strategies have been developed and applied in the identification and quantification of proteins. Among them, MS-based strategies have become the method of choice in most studies [7]. The advent of new MS instruments has improved the throughput and depth of the proteomic analysis by an order of magnitude. MS-specific resources are also quickly growing, and many efforts have been undertaken to satisfy the increasing needs for fast and reliable analysis of the proteomic data.

First of all, I will give a brief introduction about mass spectrometry and tandem mass spectrometry technology.

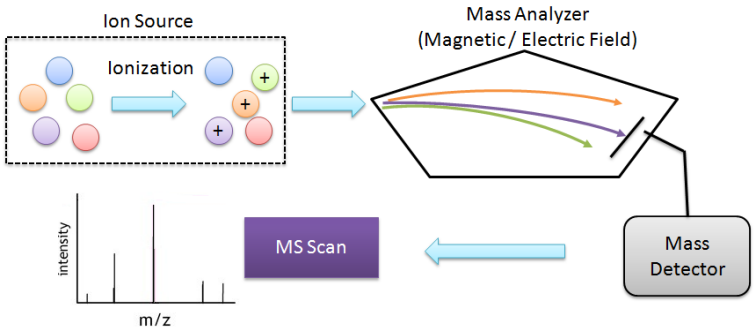


Figure 2.1: The basic components of a mass spectrometer: an ion source, a mass analyzer and a mass detector. The relative abundances of measured ions are reported in mass spectra.

2.1.2 Mass Spectrometry Technology

Mass Spectrometry

Mass spectrometry is an analytical technique that measures the mass-to-charge ratio (m/z) of charged particles. It can be used for determining the elemental composition of a sample or molecule, and for elucidating the chemical structures of molecules, such as peptides and other chemical compounds.

As shown in Figure 2.1, a mass spectrometer consists of three parts: an ion source, a mass analyzer and a detector. An ion source is used to convert analyte molecules or atoms into charged particles, which are called "*ions*". A stream of ions is bent by the magnetic and/or electric field of the mass analyzer, and only ions of a certain m/z value can reach the detector. By changing the magnetic and/or electric field, ions with varying m/z values can be measured by the instrument. A mass detector registers the number of ions at each m/z value, and the results are reported in a *mass spectrum* which shows the relative abundance of each measured ions. The mass is usually measured in Dalton (Da), which is 1/12 of the mass of a carbon atom, and is approximately the mass of a hydrogen atom.

There are two important parameters of a mass analyzer: mass resolution and mass accuracy. The mass resolution is the measure of the ability to distinguish two peaks of slightly different m/z. The mass accuracy is the ratio of the m/z measurement error to the true m/z. It is usually measured in ppm (parts per million, 10^6). There are many types of mass analyzers: quadrapole, time-of-flight (TOF), ion trap, Orbitrap, Fourier transform

ion cyclotron resonance (FT) and so on. Mass resolution and accuracy for each type of mass analyzer is shown in Table 2.1. In this thesis, the data sets are collected mainly from the high resolution FT/Orbitrap mass spectrometer, and low resolution ion trap tandem mass spectrometer.

Table 2.1: Mass resolution and accuracy of mass analyzers.

Mass Analyzer	Resolution	Accuracy
Quadrupole	2,000	100 - 1,000 ppm
Time-of-flight(TOF)	10,000 to 20,000	10 - 100 ppm
Quadrupole ion trap	20,000	10 - 100 ppm
Orbitrap	30,000 to 60,000	0.1 - 1 ppm
Fourier transform mass spectrometer (FTMS)	100,000 to 1M	0.1 - 1 ppm

Tandem Mass Spectrometry

Tandem mass spectrometry, also known as MS/MS, involves two stages of mass spectrometry, with some form of molecule fragmentation occurring between the stages. The first mass spectrometer isolates a desired target ion (charged peptide) from many entering it. This ion, also called a *precursor ion* or *parent ion*, is fragmented into product ions and sorted by the second mass analyzer. Figure 2.2 illustrates the possible fragmentation sites of a peptide. Fragment ions are labeled consecutively from the N-terminus (amino group) as a-, b- and c-ion, and also from the C-terminus (carboxyl group) as x-, y-, and z-ion. The subscript 3 of y_3-ion indicates the number of amino acid R groups this fragment ion contains. The most common and informative ions are generated by fragmentation at the amide bond between amino acids, resulting in b-ions if the charge is retained by the N-terminal part of the peptide and y-ions if the charge is retained by the C-terminal part.

There are various methods to fragment molecules in MS/MS, including collision induced dissociation (CID), electron capture dissociation (ECD), electron transfer dissociation (ETD), higher energy collisional dissociation (HCD) among others. CID is currently the most commonly used fragment method, while other methods are used to enrich certain types of ions.

The scan which measures the peptides entering the spectrometer during a fixed time interval in the first stage is called *survey scan* or *MS scan*. Subsequently, a particular

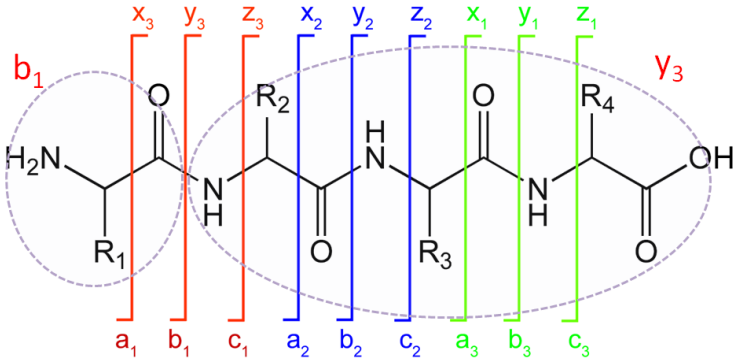


Figure 2.2: Fragmentation sites of a peptide and fragment ions produced. a-, b-, and c-type ions contain the N-terminus; x-, y-, and z- ions contain the C-terminus; R_1, R_2, R_3 and R_4 represent the side chains of amino acid residues.

peak in the MS scan is selected. The instrument will fragment the corresponding ion and measure its product ions to form an *MS/MS scan*. Usually, one MS scan is followed by one to four MS/MS scans, each targeting a different peak in the MS scan.

2.1.3 MS/MS-Based Shotgun Proteomics Strategy

A typical experimental procedure using tandem mass spectrometry to characterize a biological sample is shown in Figure 2.3. The protein complex is first purified using appropriate approaches. Isolated proteins are digested (most often using trypsin) to generate a mixture of peptides that can be identified by mass spectrometer. Liquid chromatography (LC) or high performance liquid chromatography (HPLC) is used to separate the peptide mixture before they are introduced to the ion source of the mass spectrometer. The chromatography column is located immediately in-line with the mass spectrometer, and peptides are analyzed as they elute from the column. The elute time, also called *retention time*, is different for various peptides due to their affinity and other chemical attributes, such as the mass or the charge of a peptide. The co-eluting peptides are ionized and scanned by the mass spectrometer at a particular time to obtain a survey scan (MS spectrum). Each peak in the MS spectrum ideally corresponds to a peptide. Each survey scan gives a snapshot of the peptides eluting from the LC (HPLC) column during a fixed time interval. Then, individual peptides with high abundance in the survey scans are selected as precursor ions and are fragmented using either CID or ETD fragmentation method in the tandem mass spectrometer. Finally, a large number of MS/MS spectra are produced. They are anno-

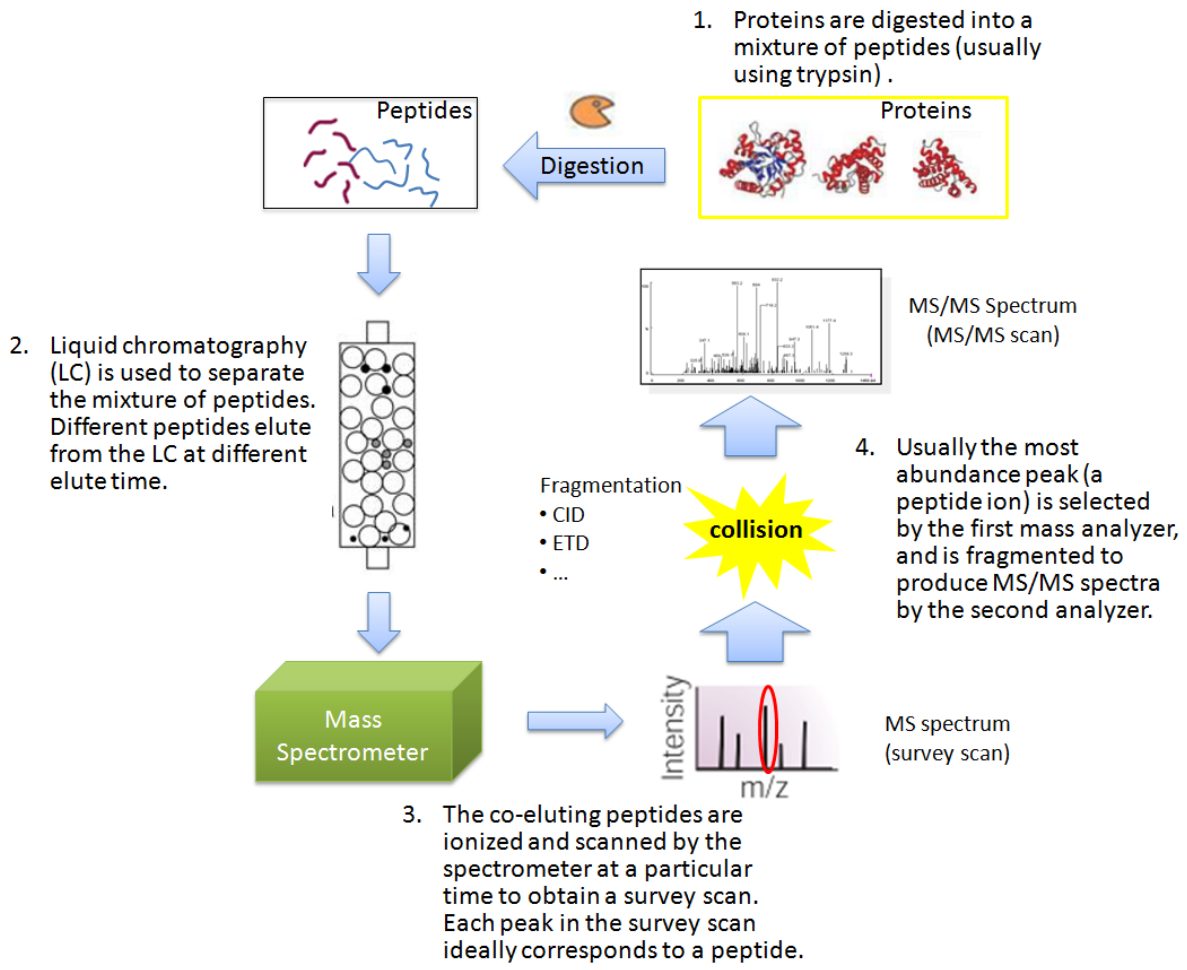


Figure 2.3: A typical experiment procedure for protein complex identification using tandem mass spectrometry [1].

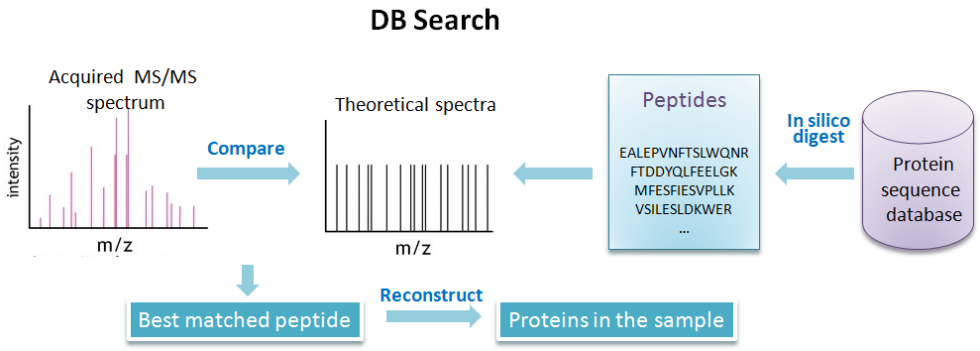


Figure 2.4: Database search strategy for peptide identification.

tated by computer-assisted software tools to indicate the peptide and protein composition in the protein complex. This is called a "bottom up" strategy in the shotgun proteomics analysis.

2.2 Peptide Sequence Assignment to MS/MS Spectra

2.2.1 Database Search

Database search remains the most reliable and widely used approach for assigning peptide sequences to MS/MS spectra. The experimental MS/MS spectra are taken as input to compare with theoretical spectra generated by peptides digested in silico from the protein sequence database (seen in Figure 2.4). A score function is used to compute the similarity between the acquired MS/MS spectrum and the theoretical spectra. Importantly, the comparisons are performed against only a set of possible peptides, filtered by applying a few criteria: the error tolerance of precursor ion, enzyme digestion constraints (allowing tryptic only or semi-tryptic), whether PTMs are allowed (if yes, the maximum number of allowed PTMs per peptide), and the fragmentation method (e.g., CID or ETD) being used. The output of the search program is the best matched peptide for each spectrum, from which we can reconstruct a list of possible proteins contained in the sample.

Many efforts have been done to develop software tools using a database search approach, such as Mascot [3], Sequest [4], X!Tandem [5], OMSSA [6] and Peaks [2]. Until today, database search is still the most widely used method for peptide, protein identification.

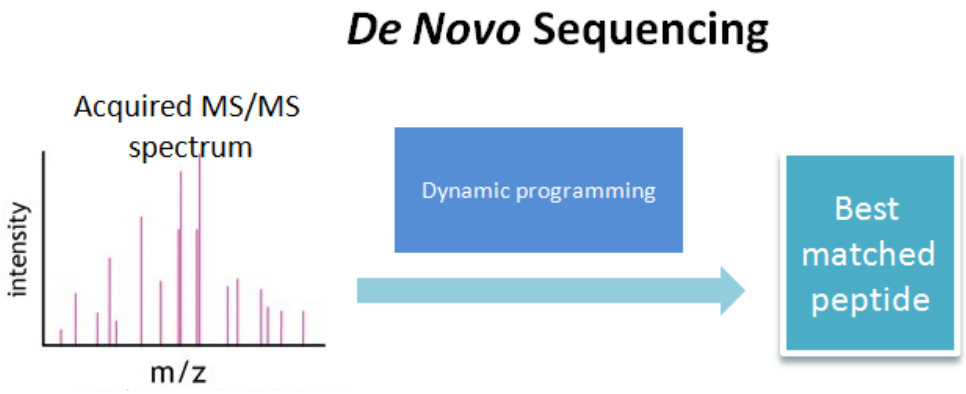


Figure 2.5: *De novo* sequencing strategy for peptide identification.

2.2.2 *De Novo* Sequencing

De novo sequencing is another approach to compute a peptide sequence, as shown in Figure 2.5. Compared with database search, *de novo* sequencing does not need to refer to any protein database, the peptide sequence is derived directly from the acquired spectrum. *De novo* sequencing is the ideal tool to identify novel sequences or when there is no sequence database available. Many *de novo* algorithms have been developed over the years, including Lutefisk [23], PEAKS [2], SHERENGA [24], PepNovo [25], etc. However, *de novo* sequencing is not practically used in large scale data analysis because it is computationally intensive and requires high quality MS/MS spectra.

2.3 Unrestricted PTM Search

A major challenge in the study of proteomics is the ubiquitous incorporation of hundreds of post-translational modifications of proteins [16]. PTMs are the chemical modifications of a protein after the translation. PTMs extend the functions of protein by attaching other biochemical functional groups such as acetate, phosphate, various lipids and carbohydrates, by changing their structures, like the formation of disulfide bridges, or by making changes in their cellular locations and dynamic interactions with other proteins. The study of PTMs will help us understand biological phenomena and the disease states involving these proteins [17].

However, a short coming of the typical database search tools is the identification of

modified peptides. Only a limited number of variable PTMs can be supported and users have to guess the possible PTMs in their data sets before the search routine is executed. As a result, a number of strategies have been recently proposed for supporting unspecified PTM search.

Many sequence tag-based tools, including the first tag-based search algorithm by Mann et al. [26], GutenTag [27], OpenSea [28] and Spider [29], have been proposed to identify inexact peptides from a sequence database. In this approach, a *de novo* sequence tag is computed from the MS/MS spectrum and used to find the approximate matches in a sequence database. The differences between the tag and a database sequence can be explained by both mutations and PTMs. The InsPecT [11], MOD^i [30] and ByOnic [31] software systems employ a hybrid approach: InsPecT uses partial *de novo* sequencing tags to perform a candidate peptides filtration in order to speed up the search; whereas the actual comparison between the MS/MS spectrum and the peptide sequence was achieved by a new dynamic programming algorithm. The algorithm automatically finds the optimal mass shifts (possible PTMs) of the amino acids to most accurately align the spectrum with the peptide. The MOD^i system applies an effective and more straightforward algorithm to compare the spectrum and the peptide. Due to speed concerns, MOD^i accepts up to 20 proteins as its sequence database, which is insufficient for the study of complex protein mixtures. The ByOnic software uses "lookup peaks" to extract candidate peptides from the database. Commercial software such as the Paragon^{TM} algorithm (Paragon) [32] and Mascot (Error Tolerant Search Mode) [33] is also available. To avoid the combinatorial explosion of the search space, Paragon uses *de novo* sequencing tags to locate "hot" areas in the protein database, where a large set of modifications are tried, while Mascot allows only one type of modification per peptide (except the specified PTMs).

2.3.1 Preliminary Knowledge About PTM and Modified Positions

In general, PTM refers to the modification that occurs after the protein translation process, both in vivo and in vitro. We can divide PTMs into two categories according to their formation periods: protein level PTMs and peptide level PTMs. Protein level PTMs refer to the modifications that have formed before the sample preparation, while peptide level PTMs refer to the modifications observed at peptides after the sample preparation. During the sample preparation, proteins are digested into a mixture of peptides, and new chemical modifications may be introduced on purpose or by unwanted reactions.

Due to the different formation periods, PTMs may occur at different modification

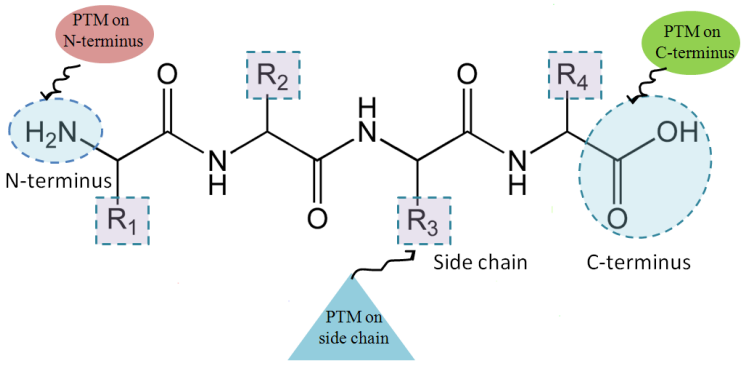


Figure 2.6: Three kinds of positions at which a PTM may occur on a protein or peptide. Circle attached on NH_2- group represents a PTM added on N-terminus. Circle on -$COOH$ group represents a PTM added on C-terminus. Triangle attached on R_3 stands for a PTM added on side chain of a residue, and R_1, R_2, R_3, and R_4 are all possible modified positions, depending on the PTM's target residue type.

positions. At protein level, there are three kinds of positions where a modification may occur: 1) side-chain of its target residues; 2) protein N-terminal amino group; 3) protein C-terminal carboxyl group. Figure 2.6 illustrates these possibilities.

Even though chemical modifications are supposed to be added to certain residues during the sample preparation, the reaction will not happen if there is already a modification occurring on the target residue. This is one of the reasons why unexpected modifications are observed to happen on cysteine (C) rather than the expected carbamidomethyl (C+57).

In "bottom-up" shotgun proteomics, identification for PTMs initially focuses on the modification at peptide level. After being identified, peptides with PTMs are then mapped to the entire protein for further study about the functions of certain modifications on proteins. Similar to protein level PTMs, there are three kinds of positions where a PTM may occur on a peptide after the protein has been subjected to proteolysis, as shown in Figure 2.6: 1) side chain of a target residue; 2) peptide N-terminal amino group; 3) peptide C-terminal carboxyl group.

For computational purpose, PTMs are modeled as two types of modification: *fixed modification* and *variable modification*, depending on the frequency of a PTM on its target residues. Fixed modification is believed to occur on the majority of its target residues. Carbamidomethyl (or carboxylation) on cysteine, which replaces a hydrogen atom of a methyl group by carbamide, is often considered as a fixed PTM. It is a side-product during the protein sample reducing process, aiming to unfold proteins by removing the disulfide

(S-S) bonds prior to mass spectrometry. PTMs that may or may not be present on its target residues, are named as variable PTMs. Most PTMs are considered as variable PTMs in the peptide identification program.

2.3.2 Error Rate Estimation for Peptide Spectrum Matches

With the improvement of PTM search algorithm, more and more modified peptide spectrum matches (PSMs) have been found, although some of them are incorrect. Incorrect identifications occur for many reasons. For example, the actual peptide sequence is not in the search database and a spurious peptide with PTMs accidently matches the spectrum well; or low-quality MS/MS spectra are used for the database search.

Regardless of the source of false identification, it is important to be able to assess which of these PSMs are correct. The raw score or a commonly used statistical measure *p-value* or *E-value* can only give a single-spectrum level confidence for each identification, but cannot provide a global measure on the whole data set. To summarize statistics for the entire collection of PSMs on a large scale data set, the most widely accepted and used statistical confidence measure is the false discovery rate (FDR). The concept of FDR was first proposed from the work of Benjamini and Hochberg in 1995 [34]. In the mass spectrometry context, FDR is defined as the expected portion of incorrectly identified PSMs among all the accepted PSMs, where an accepted PSM is the one whose score is above the threshold.

2.3.3 Target-Decoy Strategy for False Discovery Rate Estimation

Basic Concept

A simple approach to compute the false discovery rate in the MS-based proteomics is based on the target-decoy database search [35] [36] [37]. This involves doing the database search against a composite protein database which contains both the original proteins (the target part) and a portion of amino acid sequences not occurring in nature (the decoy part). The basic assumption is that the incorrect PSMs from target proteins and the decoy PSMs from decoy proteins follow the same score distribution. Therefore, the computable score distribution of the decoy PSMs, is used to predict the unknown score distribution of the incorrect target PSMs. A target-decoy method is favored in practice because of its simplicity and robustness to the effects of database size, sample quality, experimental environment and instrument types [35].

Ways to Perform Target-Decoy DB Search and Calculate FDR

There are two ways to perform a target-decoy database search:

- A single search against a database generated by concatenating the decoy database to the target database;
- Two separate searches against the target database and the decoy database;

With a given score cutoff, the FDR is calculated as:

$$FDR = \frac{D}{T} \tag{2.1}$$

Here T is the number of target PSMs with scores above the cutoff, and D is the number of decoy PSMs with scores above the cutoff.

The two separate searches may overestimate D since all MS/MS spectra are allowed to match a decoy sequences, even those correctly matched with target sequences. The single search against a concatenated target-decoy database has less bias to this problem due to the peptide competition: the true match from the target peptide likely acquires a higher score than the ones from the decoy peptides, thus only the target PSM is chosen in the identification result. In practice, the single search approach is widely accepted and used by most studies.

In some research work, FDR is calculated as $FDR = \frac{2D}{T+D}$. In this case, the peptide identification result includes both target PSMs and decoy PSMs. Considering the incorrectness of the decoy PSMs, it is more reasonable to return a result containing only target PSMs. Thus, Formula 2.1 will be used throughout the thesis.

Decoy Sequence Generation

The construction of decoy sequences should fulfill some mass-spectrum relevant properties. For example, the amino acid composition should be the same in both the decoy and target parts of the database, and the protein mass should be preserved. Furthermore, an ideal decoy database should contain at least the same number of decoy peptide sequences as the target database, allowing an a priori non-biased selection of decoy and target proteins. There are several strategies that satisfy most of these requirements [36] [37]:

1. Reverse: Reverse a protein sequence as a decoy sequence;

2. Shuffle (protein level): Each amino acid from the target protein sequence is put to a random position in the decoy protein sequence;

3. Shuffle (peptide level): The amino acids between every two consecutive digestion sites are randomly permutated, while the amino acid at each digestion site is unchanged.

The first method is straightforward and can be easily implemented, so it is used by many researchers. However, after reversing the sequence, the digestion sites of the new decoy protein are changed, causing a large number of peptides with random mass values. These peptides may not match the precursor mass of any MS/MS spectrum. The second approach does not overcome this disadvantage, either. The third method, on the other hand, does not change the mass value of any single peptide by keeping the digestion site unchanged at the peptide level. The *digestion site* (also known as *cleave cite*) refers to the cleave locations where a specified enzyme is used to digest a protein. If trypsin is used, it cleaves peptide chains mainly at the carboxyl side of the amino acids lysine (K) or arginine (R), except when either amino acid is followed by proline (P).

2.4 Two Related Problems for Peptide Identification

Besides the technical improvement in PTM search, the accuracy and speed of peptide identification can be improved in other ways. For example, assign an accurate precursor mono-isotopic mass and charge state to each MS/MS spectrum. Moreover, the preprocessing of MS/MS data could benefit all the MS data analysis software by reducing the complexity and the ambiguity of the data.

2.4.1 Preliminary Knowledge for Mass Spectrometry Data Analysis

A single mass spectrum consists of a list of ion mass-to-charge (m/z) ratios and their abundance values. For ions with positive charges, the charges arise from the addition of protons (whose mass values approximate 1 Da). Here we assume the mass of a proton $m(H^+)$ is 1 Da, and we let z be the charge state of an ion. The relation between the ion mass m and its observed m/z value mz can be calculated by using the following equation:

$$mz = \frac{m + z \cdot m(H^+)}{z} \approx \frac{m + z}{z} \tag{2.2}$$

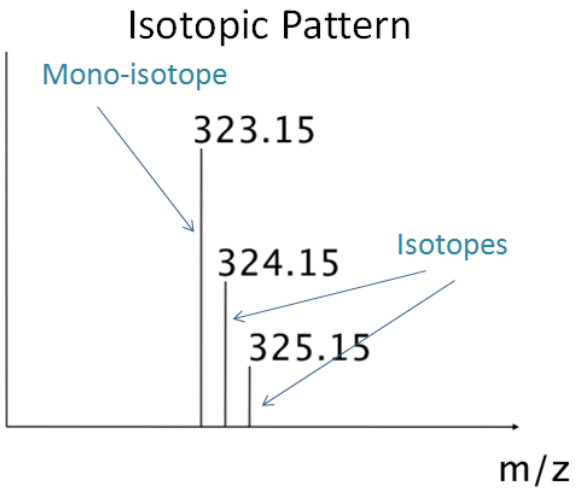


Figure 2.7: An example of an isotopic pattern. Peak with m/z 323.15 Da is the mono-isotopic peak of this isotopic envelope. Peaks with m/z 324.15 Da and 325.15 Da are two isotopes. The three peaks together compose an isotopic pattern of an unknown ion.

Due to the natural abundance of isotopes, such as C^{12} and C^{13}, H^1 and H^2, a population of the same molecular species produces a pattern that reflects the incorporation of the different isotopic contributions [38]. As a consequence, a charged ion is observed not as a single peak but as a pattern of peaks whose relative intensities and m/z values depend on the isotopic distribution of the elements they are composed of and the resolution of the instrument (see Figure 2.7). Simply selecting each observed peak as a unique ion would give rise to too many false positives. The proper way to infer an ion from mass spectrum is to group the pattern of related peaks together into an explanatory isotopic pattern, which is typically referred to as a de-isotope process. The isotopic pattern is also called an "*isotopic envelope*" or "*isotopic cluster*". The only difference between two isotopes is the number of protons. Thus, the observed inner spacing d between two adjacent isotopes can be used to calculate the charge state z of the ion (an example is given in Figure 2.8):

$$z = \frac{1}{d} \tag{2.3}$$

2.4.2 MS/MS Spectra Preprocessing

There are three general steps to preprocess MS/MS data:

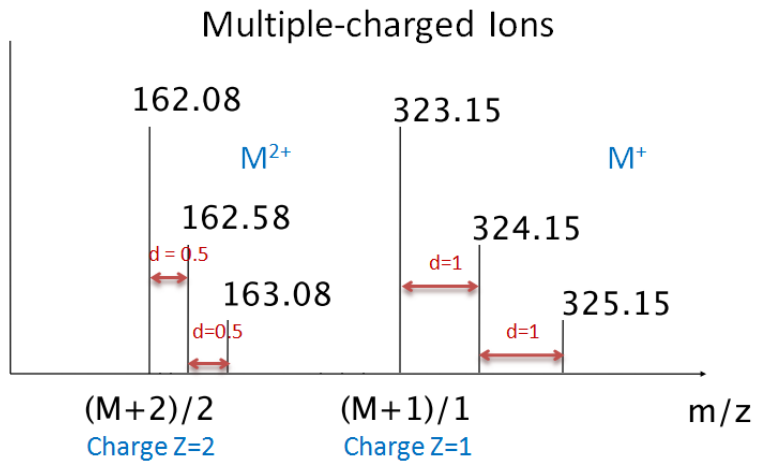


Figure 2.8: An example of different charged isotopic envelopes of the same fragmented ion. Peak list {162.08, 162.58, 163.08} is its isotope pattern with charge 2; peak list {323.15, 324.15, 325.15} is its isotope pattern with charge 1. The inner distance $d = 1/z$, where z is the charge state that equals 1 or 2.

- **Centroiding:** Due to the digitalization and measurement error of the instruments, in an unprocessed MS/MS spectrum, any peak is actually a cluster of close-by peak signals. Centroiding is the process that finds the center of this cluster, removes all the other signals except the center one, and accumulates all the heights of this cluster as the height of the center peak. After centroiding, peaks would look like the ones in Figure 2.7.

- **De-isotope:** Add all the intensities of isotopes to the mono-isotope peak, then remove all the isotopes while keeping only the mono-isotopic peak. Take the right isotopic envelope in Figure 2.7 as an example, only the peak with m/z 323.15 Da will be left with its intensity as the sum of the three isotopes.

- **Multiple Charge Deconvolution:** Combine the isotope patterns derived from the same fragmented ion but with different charge states as a unique single-charge peak. The two isotopic envelopes in Figure 2.8 are first de-isotoped, leaving only two single peaks with m/z values 162.08 Da and 323.15 Da, and the total heights of each envelope as intensities, respectively. Two peaks are then combined together: the intensity of peak 162.08 Da will be added to the peak 323.15 Da, and only the peak 323.15 Da is remained to represent their parent ion.

Chapter 3

Unrestrictive PTM search

Conventional database search tools such as Mascot [3], Sequest [4], X!tandem [5] and OMSSA [6] have been used for analyzing MS/MS spectra for about 10 years. However, the above tools are insufficient for the identification of peptides containing PTMs. In this chapter, we present an improved software tool for peptide identification with unspecified PTMs. The improvements in this software tool include (1) a default setting whereby the software considers all PTMs included in the Unimod database as variable PTMs and (2) several searching strategies are employed to reduce the search time.

3.1 Method

For the identification of modified peptides from a complex protein mixture such as the whole proteome, our computational method is designed for high-throughput MS data from a typical LC-MS/MS experiment of a protein mixture. The algorithm makes use of the high mass accuracy of the precursor ions in the MS scans; therefore, a high resolution mass spectrometer is needed for the MS scans. However, the MS/MS scans of the data can be measured with a low-resolution mass analyzer. Both data sets (human heart and yeast) used in the Section 3.2 of this chapter were measured with the LTQ Orbitrap mass spectrometer (Thermo Fisher Scientific^{TM}, Bremen, Germany). The MS and MS/MS spectra were measured with FT and ion trap, respectively.

3.1.1 Method Overview

PeaksPTM utilizes a two-pass search approach. The first pass is a traditional database search by using PEAKS software for protein identification with only a few specified PTMs. The second pass searches for modified peptides of those identified proteins, while considering all the PTM types from the Unimod library. The computational analysis consists of four major steps:

- **Protein candidate identification:** The MS/MS spectra are used to perform a traditional database search using PEAKS 5.2 software for the identification of a list of protein candidates. All identified proteins in this pass are recorded for the next pass.
- **Single-PTM peptide candidate search:** Every protein candidate is digested *in silico* to a set of peptides, and an exhaustive search is performed on these peptides to find single-PTM peptide candidates. This one-PTM-per-peptide limitation avoids the exponential growth of the search space. The identification of peptide candidates with multiple PTMs will be introduced in the final step – "multiple-PTM peptide candidate search".
- **Peptide rescoring:** All the peptide candidates are rescored by combining the peptide-spectrum matching score and the scores for two PTM-related features, the *peptide pair* and *PTM rareness*.

    1. Peptide pair: This feature examines a modified peptide candidate to determine if its base form can be independently identified from another MS/MS spectrum. For example, we see a spectrum is matched with a peptide "AATIVATSEGSLM(+15.99)GLDR". If another spectrum is identified as the peptide "AATIVATSEGSLMGLDR", which is of the same amino acid sequence but lacks of a modification "oxidation (+15.99)" on methionine (M), we call these two peptide-spectrum matches (PSMs) a peptide pair. The co-identification of the pair of modified and base forms of the same peptide increases the identification confidence.
    2. PTM rareness: A modified peptide with a rare PTM has to obtain a higher identification score in order to receive the same level of confidence as a peptide with a common PTM. This feature adjusts the score of a modified peptide candidate according to the commonality of the PTM.

The rescoring in this step is particularly important to the accuracy of identification, because the existing database search score of PEAKS does not consider any PTM-related feature. These two new features can help to evaluate the confidence of the identification of modified peptides:

- **Multiple-PTM peptide candidate search:** The common PTM types identified in the third step are used to search for modified peptides containing two or more PTMs. Approximately 10 to 20 commonly observed PTMs in the sample MS/MS data will be used to generate modified peptide candidates with more than one PTM, which largely decreases the number of candidates, but it is still a time-consuming process. The rescoring is done for all the peptides, and then the program will report the final PSM for each spectrum.

This method also controls result quality by a modified target-decoy approach, following the proposal designed for the two-pass search approach [39]. The estimated false discovery rate (FDR) would not be lower than its actual value. Moreover, we also propose a simple but effective strategy that combines the results of multiple PTM search engines to further improve the identification rate. The details of the analytical steps, the features for rescoring, the method to control the quality of results, and the consensus strategy are discussed in the following sections.

3.1.2 Protein Identification

The database search module in PEAKS 5.2 is used to identify a short list of proteins, including both target and decoy proteins from a pre-generated target-decoy protein sequence database. In this first-pass database search, only a few (one or two) PTMs are specified. These PTMs are different from the common PTMs that will be used as a feature in the second pass. They are set here only to help identify as many proteins in the mixture as possible. A protein sequence usually contains 600 to 1000 peptides, and the identification of a few of them is usually enough to identify the whole protein sequence. In general, carboxylation on cysteine is set as a fixed PTM, which means every cysteine (C) on the protein sequences will be replaced with $C + 57$; oxidation on methionine and deamidation are set as variable PTMs, meaning they may or may not occur on certain amino acids.

People usually choose a threshold to filter the obtained proteins. However, there is no reliable way so far to decide whether a threshold is properly set for all kinds of MS data sets. The setting of an inappropriately threshold may result in the failure of a search

program in identifying low-abundance protein. As a consequence, all the proteins identified by PEAKS, including both target and decoy proteins, are kept for future analysis.

For MS data collected from the LTQ Orbitrap spectrometer, the precursor mass error tolerance is set to 10 ppm, and the fragment ion mass error tolerance is set to 0.5 Da. The precursor mass and charge correction method is applied to eliminate instrument error, and details will be discussed in the next chapter.

After protein candidate identification, a reduced protein database is formed. The following analysis is performed on this reduced database.

3.1.3 Single-PTM Peptide Candidate Search

Each peptide in the reduced protein database is used as the base form to generate single-PTM peptides. Every peptide differs from the base form by only one modification, and each PTM from the Unimod database is considered. Suppose each amino acid has on average m different types of modifications in the Unimod database, then for a peptide with length k, mk single-PTM peptides will be generated on average. This is not a huge number and generates from a few hundred to a few thousand peptides depending on lengths and the amino acid compositions. Thus a brute-force algorithm is used instead of the sophisticated dynamic programming algorithm of InsPecT.

For any base form peptide, there are seven types of modified peptides containing a single-PTM:

1. One modification on the side chain of any residue, including the n-terminal residue and c-terminal residue;
2. One modification on the N-terminal amino group;
3. One modification on the C-terminal carboxyl group;
4. One modification whose occurrence involves both the amino group and side chain of the N-terminal residue;
5. One modification whose occurrence involves both the carboxyl group and side chain of the C-terminal residue;
6. Two modifications, one occurring on the amino group and the other occurring on the side chain of the N-terminal residue;

7. Two modifications, one occurring on the carboxyl group and the other occurring on the side chain of the C-terminal residue.

Please note that the last two cases actually allow two modifications in any terminus of the given base form peptide. In theory, the two-PTM-combination at one terminus is possible and more likely to happen than two PTMs on any other two positions. For example, the amino group of N-terminal asparagine (N) is modified by an N-terminal labeling reagent, followed by a deamidation on its side chain. Peptides containing these PTMs at one terminus are frequently observed in many previous studies. Therefore, we treat such a combination of two PTMs at one terminus as a "combined PTM", and take the last two cases into consideration in the identification of single-PTM peptides in PeaksPTM.

For a base form peptide and its corresponding single-PTM peptide sequences, the MS/MS spectra that match the precursor mass are selected and evaluated against the sequence by an efficient scoring function. The 512 best-scoring sequences for each spectrum are kept in memory with a priority queue during the search. After the search is finished, each sequence is further evaluated by the same peptide-spectrum matching score used in PEAKS 5.2 software. This score is a linear discriminant function (LDF) of three features: the original PEAKS score [2], the peptide length and the average score of the 512 best-scoring sequences for the spectrum. The LDF was optimized for the identification of unmodified peptides in PEAKS 5.2. We call such a score an LDF score. Only the top-scoring peptide is kept from each spectrum as our peptide candidate. Note that the candidate for a spectrum can be either a base form or a single-PTM peptide.

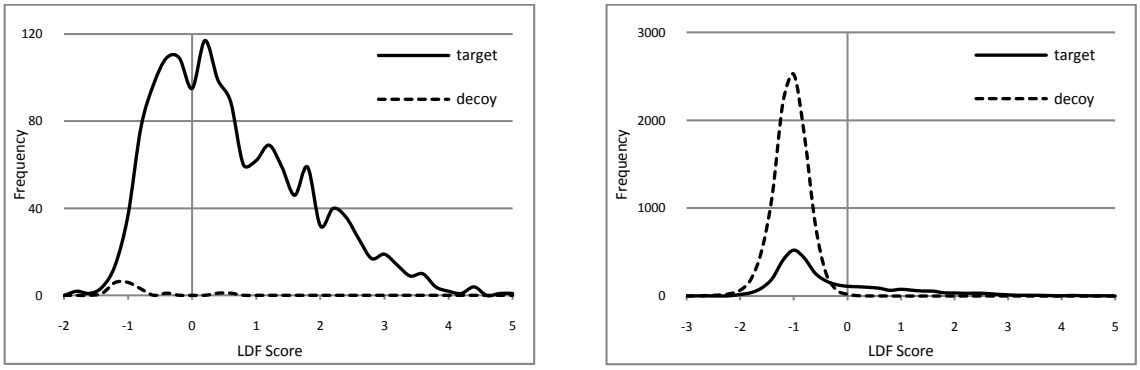
3.1.4 Peptide Rescoring

Peptide Pairs

Similar to the observations made in MS-alignment [40] and ModifiComb [41], many peptides have MS/MS spectra in the data set for both their modified and base forms. As shown in Table 3.1, MS/MS scan No.8111 and No.11626 report two peptide sequences with the same base form "AATIVATSEGSLMGLDR". Thus, it is natural to conclude that if both forms of the same peptide are independently identified from different MS/MS spectra, the identification tends to be correct. This property is illustrated in Figure 3.1(a) where the number of peptide pairs found in the target database is significantly higher than that found in the decoy database. This discovery is particularly relevant for the modified peptide candidates identified with a higher LDF score, strongly suggesting that the above conclusions are indeed correct.

Table 3.1: An example of peptide pairs in the identification results. Peptides identified for MS/MS scan No.8111 and No.11626 compose a peptide pair.

Scan #	Identified Peptide Sequence
8111	AATIVATSEGSLM(+15.99)GLDR
11174	AAVEWFDGKDFQGSK
11216	AGAYDFPSPEWDWVTPEAK
11224	EDSPSED(+28.03)PVFLR
11626	AATIVATSEGSLMGLDR
11987	AGAYDFPSPEWDW(+31.99)VTPEAK
...	.....



(a) (b)

Figure 3.1: The LDF score distributions of single-PTM peptides identified from target sequences and decoy sequences, respectively. (a) The distribution with peptide pairs, and (b) without peptide pairs. Modified peptides from the target sequences tend to have more peptide pairs than those from the decoy sequences.

Our algorithm makes use of this property by adding a reward to the modified peptide identification if the base form is independently identified from another spectrum. Please note that the identifications of the modified peptide and its base form are conducted independently. Additionally, adding the peptide-pair reward occurs only after the peptide identification is completed. Therefore, the score adjustment does not change the peptide result for any spectrum; it only affects the decision to regard the result as true or false when preparing the final report. This method is different from MS-alignment and ModifiComb which use the base form in the process of identifying the modified peptide from its spectrum. Compared to previous studies, our algorithm appears to be less sensitive – as some modified peptides may not be identifiable by their spectrum alone but only when combined with the base form. However, the specificity of our method is much improved – it is very rare that two independent identifications constitute the base and modified forms of the same peptide, unless both identifications are correct. Not only is our way of using this feature more simple, the gained specificity allows us to work more aggressively in the scoring function without creating too many false positives.

Rareness of PTMs

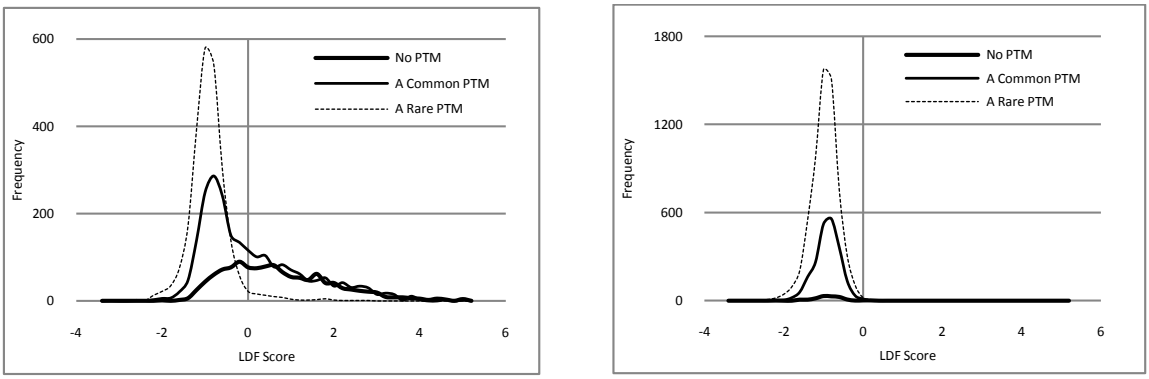
Another useful feature is the commonality of the reported PTM. A rare PTM typically demands a higher score to justify its correct identification, whereas a common PTM, such as the oxidation on methionine, is so ubiquitous that its occurrence does not require a higher score threshold than the identification of an unmodified peptide. By summarizing the common PTMs reported in previous publications [4][42], we regard the PTMs in Table 3.2 as common, while all other PTMs as rare.

Figure 3.2 shows the different score distributions of the single-PTM peptide candidates with different PTM types, from the target and decoy proteins, respectively. The decoy peptides were more randomly matched to spectra, no matter whether they had a PTM or not, or no matter which kinds of PTM they had. This phenomenon could be clearly observed from Figure 3.2(b). However, when assessing the target peptides, different frequency distributions could be seen. The distribution of target peptides with a rare PTM was still close to the distribution of the decoy peptides, while great differences were demonstrated for target peptides with a common PTM and without PTM in the high LDF score range. The feature of rareness of PTMs caused great differences in the target peptides but not the decoy peptides, suggesting a strong correlation between the PTM rareness and the identification correctness.

Since there is no quantitative measure for the frequency of each PTM type, we use N_{common_ptm} and N_{rare_ptm} to denote the number of common and rare PTMs on one peptide.

Table 3.2: The summary of 29 PTMs which are frequently reported by previous experiments.

Index	Mass	Residue	Modification name
1	-48.003372	M@C-term	Homoserine lactone
2	-29.992805	M@C-term	Homoserine
3	-18.010565	C@N-term	Dehydration
4	-18.010565	E@N-term	Pyro-glu from E
5	-17.026548	C@N-term	Loss of ammonia
6	-17.026548	Q@N-term	Pyro-glu from Q
7	-0.984016	X@C-term	Amidation
8	0.984016	N, Q	Deamidation
9	14.01565	E, D, X@C-term	Methylation
10	15.994915	W, H, M	Oxidation or Hydroxylation
11	21.981943	D, E, X@C-term	Sodium adduct
12	27.994915	X@N-term	Formylation
13	31.989828	M	Dihydroxy (Di-oxidation)
14	39.994915	C@N-term	S-carbamoylmethylcysteine cyclization (N-terminus)
15	42.010567	K, X@N-term	Acetylation
16	43.005814	K, X@N-term	Carbamylation
17	44.026215	C	Ethanolation
18	45.98772	C	Beta-methylthiolation
19	57.021465	C	Iodoacetamide derivative
20	58.005478	C	Iodoacetic acid derivative
21	71.03712	C	Acrylamide adduct
22	79.95682	Y, T, S	O-Sulfonation
23	79.96633	Y, T, S	Phosphorylation
24	99.06841	C	N-isopropylcarboxamidomethyl
25	105.057846	C	S-pyridylethylation
26	162.0528	S, T	Hexose
27	203.0794	N	N-Acetylhexosamine
28	210.19837	K, C, G@N-term	Myristoylation
29	226.07759	K, X@N-term	Biotinylation



(a) (b)

Figure 3.2: The LDF score distributions of the peptide candidates identified with no PTM, a common PTM, and a rare PTM, from (a) the target and (b) decoy proteins.

For both PTM types, penalties are obtained from training. The penalty for any modified peptide is the sum of the penalties of PTMs.

Weighted Sum Score

Our final score for a modified peptide candidate is a linear combination of four features: the PEAKS LDF score, the number of common PTMs, the number of rare PTMs, and the existence of a peptide pair. More specifically, the score is defined by

$$S_{ldf} + c_1 \cdot E_{peptide\_pair} - c_2 \cdot N_{common\_ptm} - c_3 \cdot N_{rare\_ptm} \tag{3.1}$$

where $E_{peptide\_pair}$ is set to 1 if there is a peptide pair, and 0 if there is no such a pair. The coefficients c_1, c_2, and c_3 are obtained by training.

One obstacle for parameter training is to find a data set with correct modification annotations. Large-scale manual annotation is impractical. Simulated data sets were used in previous research, but the introduced false negative was difficult to evaluate [11]. Alternatively, we trained the coefficients by maximizing the number of identifications at 1% FDR, estimated with a target-decoy approach. Great care was taken to avoid the possibility of over fitting: an independent data set (the yeast data set) was used as training data and the generated coefficients were used to test the human heart data set.

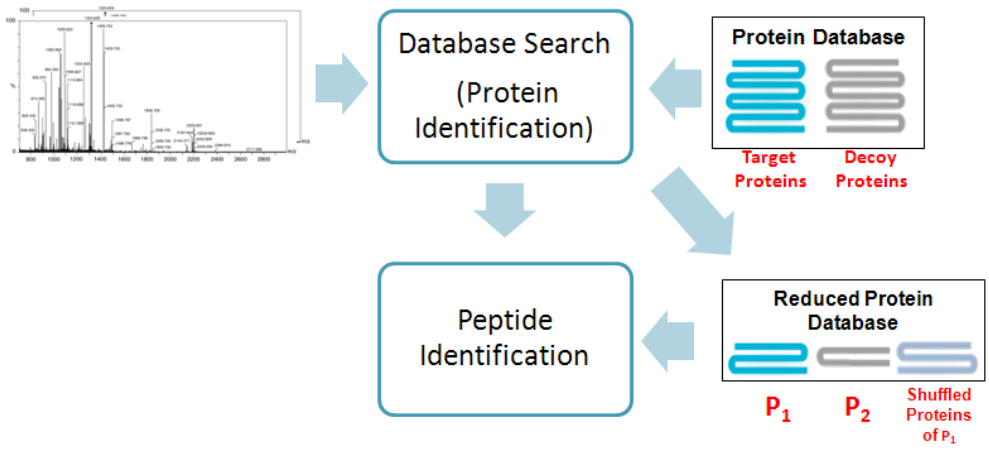


Figure 3.3: The modified target-decoy strategy used in PeaksPTM. P_1 is the set of proteins coming from the target database, while P_2 is from the decoy database.

3.1.5 Estimation of the False Discovery Rate

Estimation of FDR By Using A Modified Target-Decoy DB Strategy

Researchers have some concern when the traditional target-decoy strategy is applied to the two-pass approach. Although there are equal numbers of target and decoy protein sequences in the first pass, there are always more target sequence than decoy sequences chosen for the second pass. Even worse, the small decoy database might not be able to reflect the distribution of falsely identified PSMs on the target sequences.

Designed for a two-pass approach, a modified target-decoy strategy was proposed recently [39], and it never underestimated the FDR. In the first pass, it still uses a target protein database concatenated with its decoy version to determine the possible proteins. A reduced database, which includes the proteins from the target database in the result of the first round search (P_1), the proteins from the decoy database (P_2), and the newly shuffled proteins of P_1, is then searched in the second round. This method is only slightly biased against target peptides, and the estimated FDR will not be lower than its actual value. PeaksPTM adopts this approach in the FDR control, as shown in Figure 3.3.

We use the following method to calculate the FDR: suppose there are D identifications from the decoy proteins and T identifications from the target proteins, the FDR after removing the decoy hits is calculated as D/T.

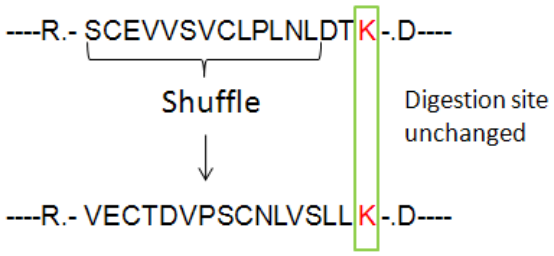


Figure 3.4: An example of the peptide level shuffled approach used in PeaksPTM.

Decoy Protein Sequence Generation

In PeaksPTM, the decoy sequences are created by using the peptide level shuffling approach. As discussed earlier, this approach preserves the peptide precursor mass of the decoy peptides. An example of shuffling a peptide is shown in Figure 3.4, with trypsin as the enzyme. A peptide "SCEVVSVCLPLNLDTK" matches a spectrum, and its shuffled version, say "VECTDVPSCNLVSLLK", with the same mass value will still match the precursor mass of that spectrum. Therefore, the shuffled peptide can be used to evaluate the confidence of identification of the target peptide. If the shuffled peptide achieves a similar or even a higher score, it shows that the PSM of the target peptide "SCEVVSVCLPLNLDTK" is a random match; otherwise, it is likely a correct match with high confidence.

Additionally, two details are added in the shuffling approach:

(i) If a decoy peptide has the same sequence as a target peptide, re-shuffling is done;

(ii) If the length of a digested target peptide is shorter than the 5, it will be combined with its neighbor(s) and shuffled.

3.2 Experiments and Results

Our software was compared with Mascot (Mascot 2.3, Error Tolerant Search Mode) [33], Paragon (ProteinPilot software 4.0.8085, Paragon Algorithm: 4.0.0.0, 148083, trial version) [32], and InsPecT [11] (release 20101012) with an MS/MS data set obtained from the human heart tissue. We also compared our software with MOD^i [30].

Table 3.3: The numbers of identified peptides with 1% FDR under different settings of training and testing data sets.

	Yeast (training)	Human Heart (training)
Yeast (testing)	4,286	4,219
Human Heart (testing)	2,410	2,447

3.2.1 Data Sets

Two data sets, which were downloaded from an online proteomics database [43], are used in our experiments. The sample preparation processes are described as follows:

Human heart. Heart tissue was homogenized with a Dounce homogenizer. The proteins were reduced with DTT and alkylated by iodoacetamide, then digested by trypsin overnight. The peptide mixture was separated via SurveyorT LC equipped with MicroAST autosampler (Thermo Fisher Scientific^{TM}, Bremen, Germany) using a reversed phase analytical column. The data was collected with an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific^{TM}, Bremen, Germany) consisting of 11,207 MS spectra and 15,117 MS/MS spectra.

Yeast. The yeast data set is on a fraction of Lys-C digest of a yeast lysate. It contains 5,136 MS spectra and 12,366 MS/MS spectra measured using an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific^{T}M, Bremen, Germany).

3.2.2 Cross-Training on Two Data Sets

The independent yeast data set was used to train the coefficients mentioned in Section 3.1.4 for the final score calculation. This strategy helps to avoid the over fitting problem caused by training on the same or related-species data set.

We also verified the performance by a cross-training strategy (see Table 3.3). The parameters trained on one data set were tested on the other data set. If the parameters trained on yeast data set were tested on the human heart data set, the number of identifications at 1% FDR decreased from 4,286 to 4,219. If the parameters trained on the human heart data set were tested on the yeast data set, the number of identifications decreased from 2,447 to 2,410. The cross-training results illustrate that using the same training and testing data only produces slightly better results than using different training and testing data, indicating that the over fitting problem in our method is negligible.

3.2.3 Comparison Between Multiple Search Engines

PeaksPTM was compared with Mascot (Error Tolerant Search Mode), Paragon and InsPecT to evaluate its performance. Since Mascot and Paragon have their own first round search functions, the IPI Human (v3.75) database concatenated with its shuffled version, was used as the search database. The corresponding FDRs were calculated using the standard target-decoy approach [44][36]. PeaksPTM used the same target-decoy database to find 1,349 target and 773 decoy proteins; 1,349 additional decoy proteins were then added. 3,471 proteins in total were used in the second round search. Since, in blind search mode, InsPecT could not finish the whole IPI human database, it was applied on a short list of 2,030 proteins found by PEAKS 5.2 software (regardless of their scores). This pre-selected protein list should be a superset of the high abundance proteins in the sample. The same numbers of shuffled decoy protein sequences were searched together to determine the FDR.

For PeaksPTM and Mascot, the precursor and fragment ion error tolerances were set to 10 ppm and 0.5 Da, respectively. The maximum variable modification number per peptide was set to 1 in PeaksPTM. For Paragon, we chose trypsin as the enzyme, Orbitrap/FT MS (1-3ppm) LTQ MS/MS as the instrument setting, biological modifications as the modification setting, and the thorough search mode (in contrast to the rapid search mode). For InsPecT, trypsin was designated, blind search was turned on and the variable modification number was set to 1. The 15,117 MS/MS spectra were split in two approximately equal batches for InsPecT to run in parallel on two computing cores of an Intel® Core^{TM} i7 CPU, 2.80GHz. InsPecT used 21 CPU hours in total. On the same computer utilizing two computing cores, PeaksPTM, Paragon and Mascot all finished the analysis in approximately an hour.

An MS/MS spectrum with its identified peptide is called a peptide-spectrum match (PSM), and if this identified peptide is modified, it is called a modified PSM. Figure 3.5 shows the performances of the four software packages on the identification of modified PSMs. At 1% FDR, PeaksPTM reported 2,410 PSMs, 1,394 of which were modified PSMs; Mascot reported 1,355 PSMs and 743 modified PSMs, Paragon reported 1,972 PSMs and 1,029 modified PSMs, and InsPecT reported 1,133 PSMs and 521 modified PSMs. Even using a more strict FDR estimation than the other three engines, PeaksPTM still performs significantly better than its competitors.

We further investigated the composition of the reported modified PSMs by PeaksPTM in Figure 3.6. Among the 1,394 modified PSMs reported by PeaksPTM with $\leq 1\%$ FDR, 785 (56.3%) were supported by at least one other search engine with high confidence (with FDR $\leq 1\%$), and 449 (32.2%) were supported by at least one other search engine regardless of the confidence. Since it is rare for two engines to falsely identify the same modified PSMs,

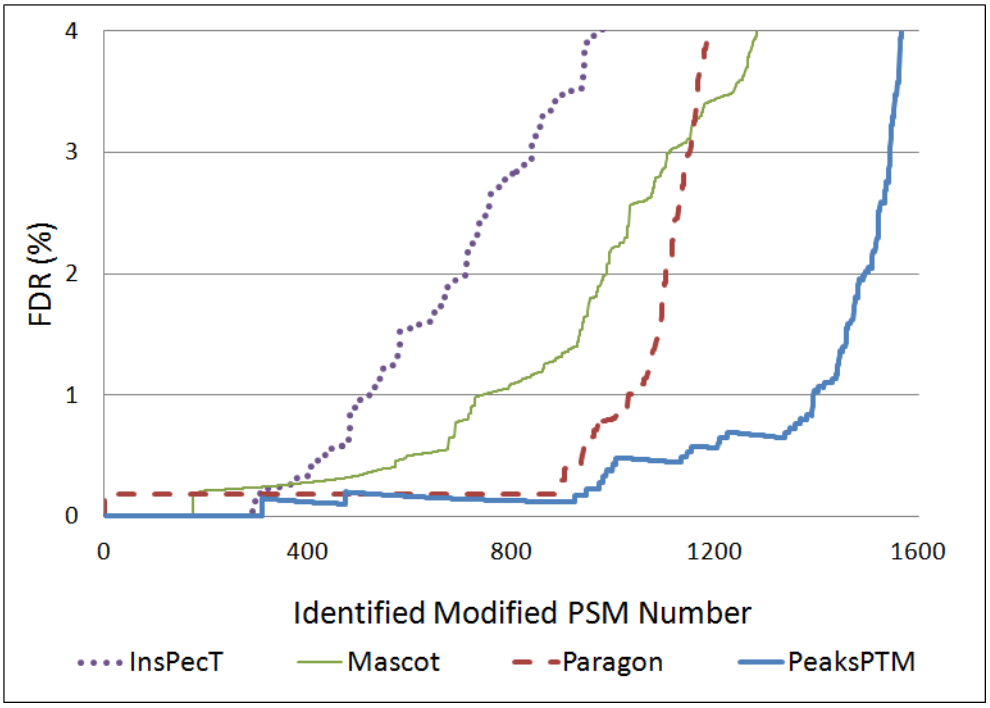


Figure 3.5: The comparison of reported modified peptides by InsPecT, Mascot, Paragon and PeaksPTM. The curves show the relation between the estimated FDR and the number of results reported.

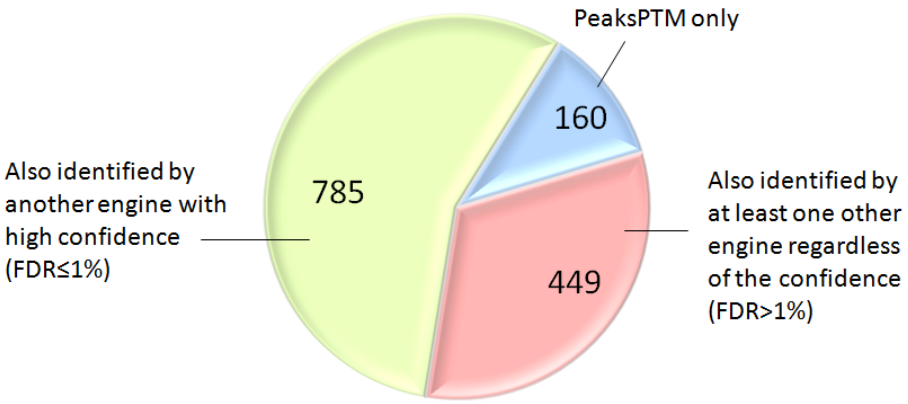


Figure 3.6: A large portion of PeaksPTM's high confidence (FDR $\leq$ 1%) modified PSMs are also identified by at least one other engine, either with high or low confidence.

these consensus identifications are of high confidence.

3.2.4 Comparison with MOD^i

MOD^i can only identify peptide sequences from a small database containing at most twenty proteins. Therefore, ten highest-scoring non-homologous proteins (out of the 1,349 target proteins from the first round search using PEAKS 5.2) and their shuffled versions were combined as the reduced protein database for MOD^i. All the modifications provided by the MOD^i web server were chosen as variable modifications, and its default setting for modified mass range ($-150 \sim 250$ Da) was used. The negative range of mass values represents the decrease of the mass of a target residue when it is modified by removing some of its chemical groups. For example, dehydration on a serine (S) will decrease its mass value by 18.01 Da. In contrast, the positive range of mass values represents the increase of the mass of a target residue when it is modified by attaching some chemical groups.

The InsPecT blind search can also be used as a second round PTM search tool, which accepts a reduced protein list generated by any standard database search. Because of this capability, InsPecT was also added to the comparison with MOD^i. For a fair comparison InsPecT and PeaksPTM were both used to search the same reduced protein database as MOD^i.

Figure 3.7 shows the comparison of three software tools. PeaksPTM still performed best in terms of finding modified PSMs. We warn researchers that because of the small size of the target and decoy protein lists, the FDR curves can only be used for the purpose

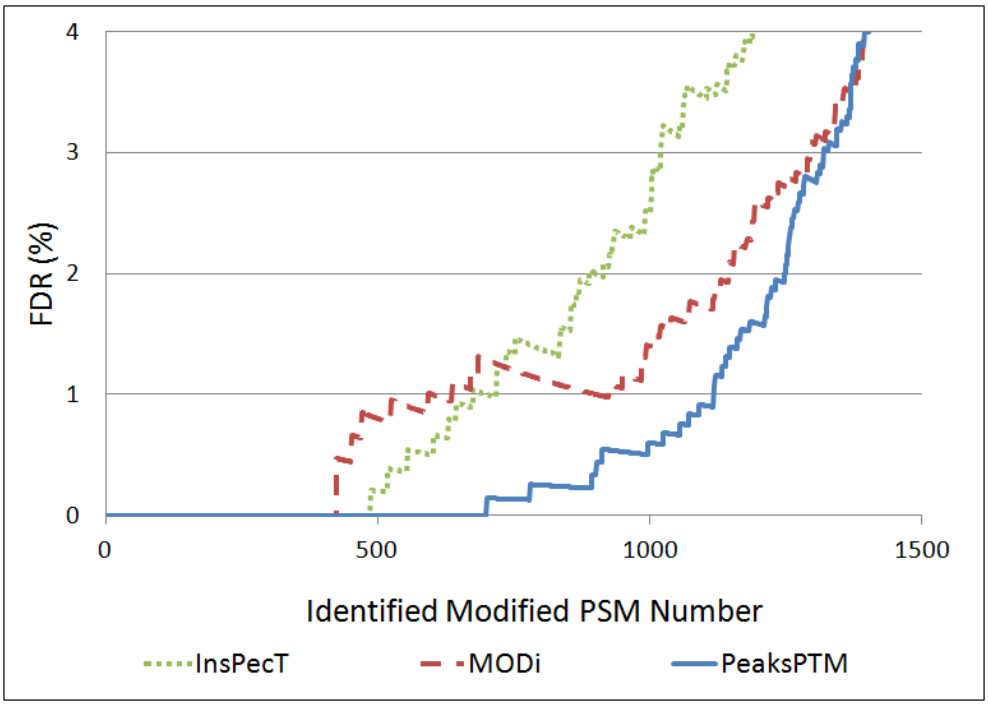


Figure 3.7: The comparison of PeaksPTM, MOD^i and InsPecT on the reduced database with 10 target + 10 decoy proteins. The curves show the relation between the estimated FDR and the number of results reported.

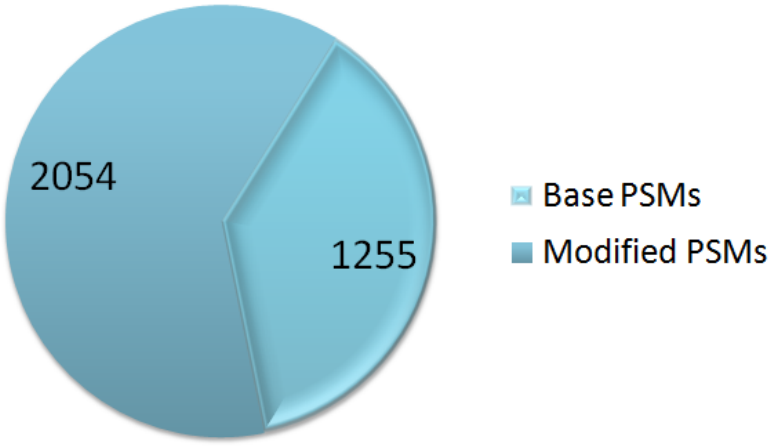


Figure 3.8: The consensus result of identified PSMs by using consensus strategy on human heart data set.

of comparing these three tools, but may not accurately reflect the real FDR values of the identifications. Additionally, the relative performances from such a small database may be very different from those of a large database.

3.2.5 Consensus Strategy and Analysis

A consensus rule that combines the identifications from four search engines: PeaksPTM, Mascot (Error Tolerant Search Mode), Paragon^{TM} and InsPecT, is adopted to improve modified peptide identification. A PSM that is identified by more than one search engine or by only one search engine with FDR less than 0.8% is considered as a confident identification.

Using this consensus strategy, 3,220 PSMs were reported, 1,965 of which were modified PSMs and 1,255 base PSMs (see Figure 3.8). This consensus strategy results in a 40% increase in the identification rate of high-confidence modified PSMs compared with any single search engine.

Figure 3.9 illustrates the composition of 1,965 modified PSMs contributed by four search engines. Two modified peptides identified by different engines from the same spectrum are regarded the same if their base forms, number of modifications and modification mass shifts are the same. Note that the modification site is insignificant in this study. This Venn diagram indicates that a large number (871) of modified PSMs were identified confidently

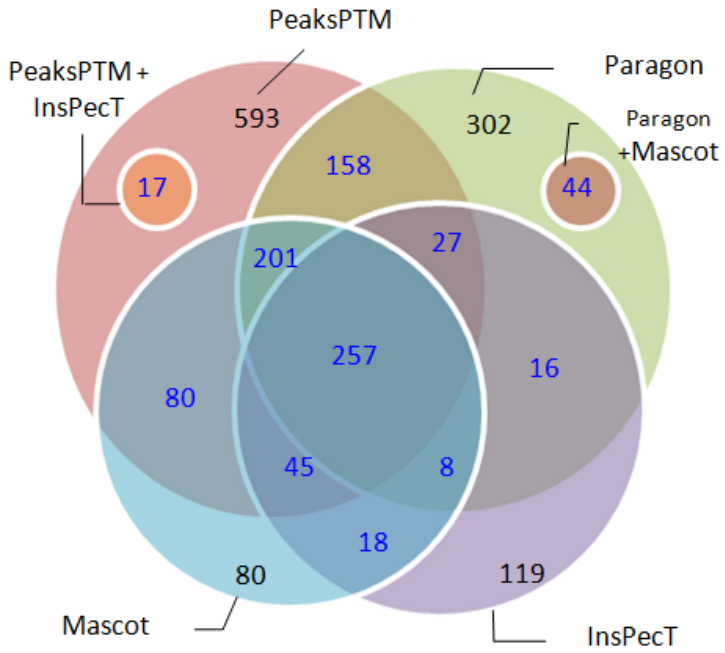


Figure 3.9: The Venn diagram shows the composition of confidently identified modified PSMs by the four search engines using a consensus strategy.

by two or more engines independently. This is over 35% of all PSMs (modified or not) identified by any single search engine alone: PeaksPTM, Mascot or Paragon could report at most 2,400 PSMs (modified or not) at 1% FDR.

The large number of highly confident PSMs confirms the belief that the inefficiency in modified peptide identification is one of the major factors for the low characterization rate of the MS/MS spectra in a data set [11] and the low identification rate of the modified peptides [16].

3.2.6 Summary of Identified PTMs

Table 3.4 summarizes the frequent PTMs identified by PeaksPTM with 1% FDR from the human heart data set. The same modified peptide, identified from multiple spectra, is only counted once in this section. There are 906 unique modified peptides identified by PeaksPTM. Oxidation is the most common PTM, occurring on 200 peptides. The utilization of the high resolution MS data enables PeaksPTM to identify modifications with small Δm, such as deamidation. But it is still possible that a PTM's name is mistakenly replaced by another PTM name with similar Δm.

3.3 Discussion

In this chapter we have discussed our improved software tool, PeaksPTM, used for identifying peptide sequences with unspecified modifications. To increase the confidence of the identification of a modified peptide, the algorithm utilizes two features: peptide pair and PTM rareness, to improve the unrestricted PTM search. Considering both features, the peptide pair seems to be most important: 86.6% of the modified PSMs confidently identified by PeaksPTM have peptide pairs. Compared to the PEAKS 5.2 LDF score alone, adding the peptide pair feature and the PTM rareness feature could identify 608 (35.9%) and 156 (9.2%) more modified PSMs at 1% FDR, respectively. Adding both features improved 717 (42.4%) identified modified PSMs.

The experimental results show that at the same FDR, our software significantly outperforms three other major search engines: Mascot, Paragon and InsPecT, in terms of the number of modified PSMs identified. Furthermore, results from multiple search engines confirmed over 871 highly confident modified PSMs, which is over 35% of the reported PSMs by any single search engine. This confirms the evidence in the literature that inefficiencies in modified peptide identification is one of the major factors for the low charac-

Table 3.4: The numbers of unique modified peptides containing the most common PTMs in the human heart data set.

Mass (Da)	Residues	Modification	PeaksPTM
-18.01	S, T, D	Dehydration	10, 6, 8
-18.01	E@N-term	Pyro-glu from E	12
-17.03	N	Loss of ammonia	8
-17.03	Q@N-term	Pyro-glu from Q	18
-2.02	S, T, Y	2-amino-3-oxo-butanoic_acid	6, 4, 3
0.98	N, Q, R	Deamidation	61, 39, 3
13.98	P	Proline oxidation to pyroglutamic acid	4
14.02	E, D, S	Methylation	84, 11, 5
15.00	N, Q	Deamidation followed by a methylation	6, 7
15.99	M, Y, F,W, H, P, N, K	Oxidation or Hydroxylation	99, 28, 25, 17, 11, 9, 6, 5
27.99	S, K, T, X@N-term	Formylation	24, 6, 8, 15
28.03	E, D	Ethylation	41, 7
31.99	M, W, P	Dioxidation	23, 13, 10
42.01	S, X@N-term	Acetylation	3, 4
43.99	W, D	Carboxylation	9, 1
47.98	C	Cysteine oxidation to cysteic acid	15
57.02	C, K, H	Carbamidomethylation	21, 3, 2
79.97	S	Phosphorylation	4

terization rate of the MS/MS spectra in a data set and the low identification rate of the modified peptides.

The second round of PeaksPTM is designed for searching modified peptides with multiple PTMs. However, when increasing the maximum allowed PTM number to 2, we only observed 32 new identified high confident modified PSMs at 1 % FDR, while the running time increased up to 3 hours. This result demonstrates (a) the human heart data set may not contain few heavily modified peptides (b) the time spent on searching the peptides with multiple PTMs is not negligible, even if only 20 variable PTMs were considered in the second round. As a consequence, we only compared the peptides with at most one PTM in our result with the identifications of other search engines. Further improvement on the efficiency of searching peptide with multiple PTMs would be a major goal in the next version of PeaksPTM.

We note that PeaksPTM is not a blind-search engine like InsPecT, which also attempts to find novel PTM types that were previously unknown. However, being able to use all PTM types in the Unimod database will be sufficient for most proteomics research today. In our experiment, InsPecT was able to identify only one modification with mass shift that did not match any PTM type in the Unimod database. Such identification definitely deserves an expert's careful examination before it is added to the Unimod PTM database. As the experimental results show, such identification of novel PTMs also decreases the level of performance on known PTMs. Therefore, we recommend that researchers choose different tools according to their specific applications. Another note is that the target-decoy FDR control method widely used today (and used in PeaksPTM) can only control the peptide sequence, but not the modification site inside the sequence. Consequently, all the FDRs reported in this chapter consider the correctness of the modified peptide sequence and the Δm of the modifications, but cannot ensure the correctness of the modification sites reported by those software tools.

In an earlier version of the PeaksPTM software, another scoring feature, the precursor pair, was used. For each modified peptide candidate, the precursor m/z and retention time of the base form could be predicted. If a significant peak was observed at the predicted location in the MS scans of the data, it was likely caused by the base form of the peptide. Thus, the identification confidence of the candidate increased. However, after an amendment to the peptide pair feature, we found the contribution of the precursor feature in the early version disappeared in the new version. As a result, the precursor pair feature has been removed from the PeaksPTM software reported here. However, it is likely that this feature may still be useful under certain experimental settings where not all base forms of the modified peptides are fragmented in the mass spectrometer to produce MS/MS spectra.

The PeaksPTM software is freely accessible at:
http://www-novo.cs.uwaterloo.ca:8080/PeaksPTM/.

Chapter 4

Two Related Problems for Peptide Identification

It is always hard to avoid errors and ambiguities in the data collection from biological samples. Therefore, any data analysis tool would benefit from the correction of these errors and the removal of the ambiguities. For MS/MS data, the accurate information in the data, such as the precursor ion, real ion signals (among the noise signals), etc., is essential for a better interpretation of the spectra, and can be achieved through effective data preprocessing. In this chapter, we will study 1) the precursor mono-isotopic mass and charge correction problem and 2) the MS/MS preprocess problem.

4.1 Precursor Mono-Isotopic Mass and Charge Correction Problem

The mono-isotopic precursor ion mass is crucial for most existing software to identify a peptide from its MS/MS spectrum. High-resolution MS/MS instruments promise to significantly enhance proteomics analysis by providing smaller mass error tolerance for both the precursor and the fragment ions. The Thermo Fisher LTQ-Orbitrap instrument is among the most popular high-resolution instruments. However, very often the precursor mass reported by these instruments differs by one or more isotopes from its correct mass value. This would cause the software analysis to fail unless, contrary to the nature of high-resolution experiments, a bigger mass error tolerance is used. An example is shown in Figure 4.1. The bottom panel of this figure is the given MS/MS spectrum with precursor

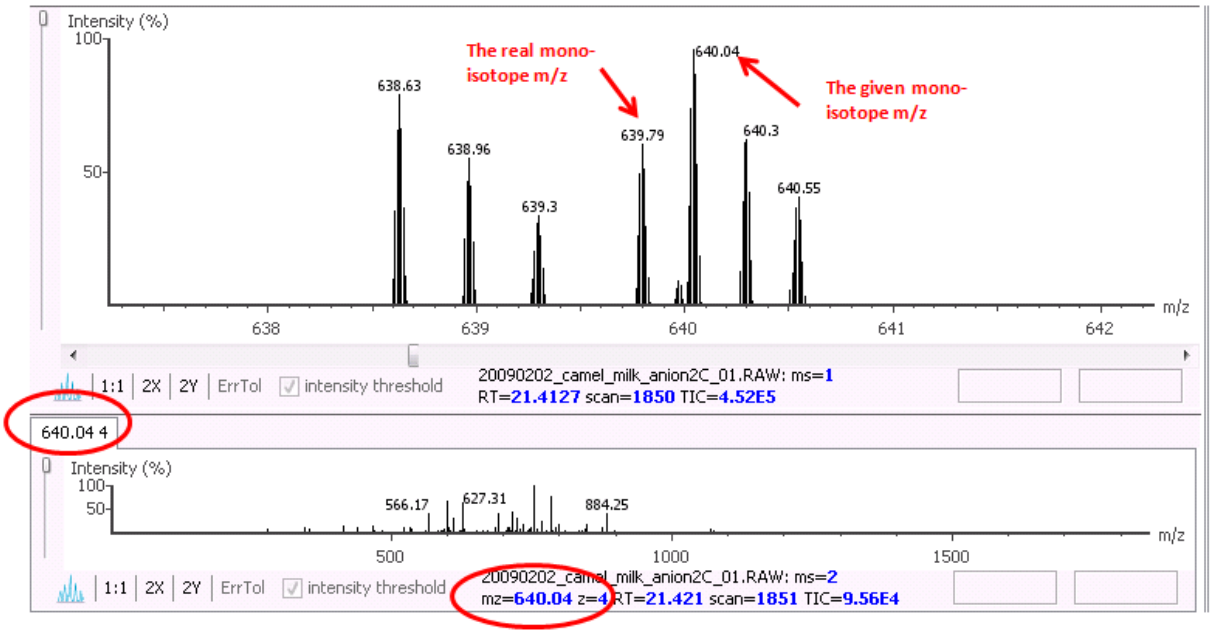


Figure 4.1: An example of precursor mass shift.

ion m/z 640.04 Da and charge state 4. If we check its parent MS spectrum, shown in the top panel of the figure, we can see that real mono-isotope of its precursor ion should be the peak with m/z 639.79 Da. This is an example of one isotope offset shift from the given precursor m/z value. In the following section, we will propose an algorithm that automatically determines the correct charge state and mono-isotopic mass of the precursor ion from high-resolution MS data.

4.1.1 Method Overview

In the parent MS scan of a given MS/MS spectrum, we can locate the m/z region of precursor ion (peptide) that gives rise to this MS/MS spectrum. Analyzing the peak shapes in this region, our algorithm is able to distinguish the isotope peaks generated by the precursor ion, and gives the correct mono-isotope mass value. This algorithm involves two steps:

- **Candidate Generation:** Based on the reported m/z value m of the precursor ion, the algorithm collects all peaks within an m/z window, ranging from $m - 5$ Da to

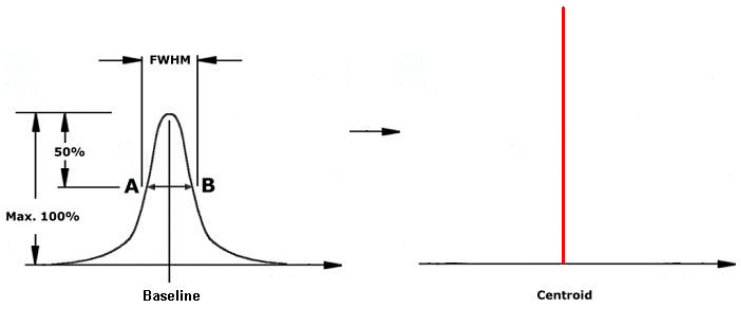


Figure 4.2: Peak centroiding.

$m + 5$ Da. Any possible isotopic envelopes within this window are extracted and a candidate list is built. The only constraint for these candidates is that their mono-isotope m/z value should be between $m - 0.001m$ and $m + 0.001m$.

- **Candidate Evaluation:** For each candidate (isotopic envelope) in the list, the theoretical isotope distribution of an average peptide with the same mass value is computed, and compared with the observed peaks in the spectrum. The candidate with the least isotope distribution fitting error is reported as the result.

4.1.2 Candidate Generation

Peak Preprocess

Peaks selected from the parent spectrum window are required to be greater than both a user-defined signal-to-noise threshold and a background noise level. The background noise is calculated as the average intensity of noise peaks which are lower than: 1) 10% of the height of the local highest peak and 2) 80% peaks in a nearby region.

In an unprocessed mass spectrum, any peak is actually a cluster of peak signals. Before any data analysis, we need to centroid these selected peaks first. As illustrated in Figure 4.2, the highest peak signal in such a cluster is considered to be a baseline. The width of the peak is taken as the full-width half-maximum (FWHM) of the curve. The intensities of peaks within this width are summed up as the intensity of the new centroided peak.

After removing the noise and peak centroiding, a peak list PL is obtained from the parent spectrum window: $PL = \{P_0, P_1, \ldots, P_n\}$, where P_n is last peak in this window. The window size is from m/z value $m - 5$ Da to $m + 5$ Da, where m is the precursor m/z reported by the instrument, and the 5 Da boundary limit is an empirical value. The m/z

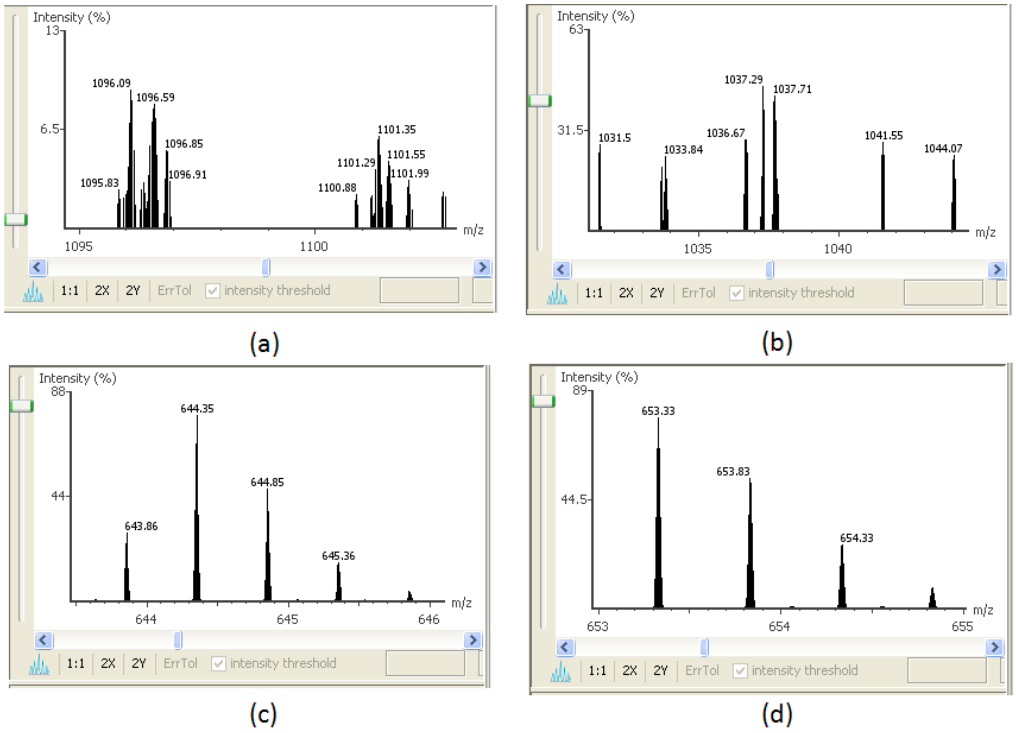


Figure 4.3: Examples of noise peaks and isotopic peaks. (a), (b) and (c) show spurious peak groups, and (d) gives an example of real isotopic envelope.

range of [-5, +5] Da is wide enough to include the entire isotopic envelope of the precursor ion.

The next step is to find every possible isotopic envelope candidate E_i in this window.

Theoretical Isotopic Abundance Distribution

Different from noise peaks which randomly occur in a spectrum, a real ion can generate an isotopic envelope with *good shape features*: peak height, equal intra peak distances, similar peak width, and relative intensities of peaks in the cluster. This is because every peak in an isotopic envelope comes from the same molecule compound with different combinations of isotopes. Lacking any of these "good shape" features, a group of peaks probably will not be considered as an isotopic envelope. Figure 4.3 gives some examples of spurious peak groups in (a), (b) and (c), as well as an example of an actual isotopic envelope in (d).

Table 4.1: The look-up table for the prediction of isotopic abundance distribution. The mono-isotopic masses in this table range from 50 Da to 75,000 Da.

mass (Da)	most abundance peak index	isotope array size	intensity						
			1^{st}	2^{nd}	3^{rd}	...	7^{th}	...	15^{th}
50	1	1	100.0						
100	1	2	100.0	6.070					
...	...	...	...						
600	1	3	100.0	33.62	6.93				
650	1	3	100.0	36.18	8.09				
...	...	...	...						
13,450	7	15	7.17	18.6	36.76	...	100.0	...	5.62
13,500	7	15	7.03	18.3	36.3	...	100.0	...	5.85
...	...	...	...						

Some of the high intensity noise peaks in (a) and (b) are narrower or wider than others, and the intra distances among them do not follow any $1/z$ pattern. The peak group in (c) is also not an isotopic envelope because its peak abundance distribution is quite different from the natural relative abundance distribution of isotopes. As we know, all protein (or peptide) molecules consist of five basic chemical elements: hydrogen, carbon, oxygen, nitrogen, and sulphur. The proportion of each element in a molecule does not significantly fluctuate with the change of molecule mass value, and the isotope abundance is mostly determined by the number of C^{13} atoms. If an ion with mono-isotopic m/z 643.86 Da and charge state 2, its mass value is 1285.72 Da. For any ion with mass value approximate 1300 Da, the height of its first isotope cannot exceed the height of its mono-isotope. Figure 4.3(d) shows an actual isotope abundance distribution derived from an ion with mass value approximate 1300 Da, compared to the one with 1285.72 Da in (c) . The difference in abundances leaves us to believe that the peak group in (c) is more likely derived from at least two overlapped ions with similar mass value: one starting from mono-isotope peak m/z 643.86 Da, and the other one from mono-isotope peak m/z 644.35 Da, and both of them are charged 2.

An algorithm, called *averagine*, is widely used to predict the isotopic abundance distribution of a molecule when its mass value is given. This algorithm assumes that every molecule can be divided into several average molecules, which have the molecular formula $C_{4.9384}\ H_{7.7583}\ N_{1.3577}\ O_{1.4773}\ S_{0.0417}$ and an average molecular mass of 111.1254 Da. Given

a molecular mass, the algorithm computes the number of average molecules involved by dividing the given molecular mass by the average molecular mass. By scaling the molecular formula of the average molecule, the algorithm computes a standard molecular formula for the given molecule. From the standard molecular formula, the algorithm can compute the isotopic abundance distribution of the given molecular mass [45].

To minimize the calculation of theoretical isotopic peak abundance, we adopt the look-up table used in the THRASH algorithm [46]. In the look-up table, masses range from 50 to 75,000 Da with intervals of 50 Da, and indexing is given for positioning the most abundant peak, as shown in Table 4.1. Due to the small interval, no interpolation is needed for the mass within intervals. The closest mass value in the look-up table is used to obtain the isotopic peak intensity distribution. For example, consider an ion with mass value 605.3 Da and intensity 6000. If checking the entry of 600 Da in the look-up table, we can obtain the relative intensities of isotopes is 100%, 33.62% and 6.93%, respectively. The theoretical isotopic intensity for this ion should be: 6000, 2017.2 ($= 6000 \times \frac{33.62}{100}$) and 415.8 ($= 6000 \times \frac{6.93}{100}$). In the look-up table, we can also see the increments in the relative abundance of isotopic peaks with the increase of the molecule mass value. When the mass value attains a value of 13,450 Da, the most abundance peak in an isotopic envelope is no longer its mono-isotope, but the 7^{th} isotopic peak.

Isotopic Envelope Selection

As mentioned earlier, we extract a peak list PL within a window from the parent spectrum. The m/z of the mono-isotopic peak of each isotopic envelope candidate is required within a range from $m - 0.001m$ Da to $m + 0.001m$ Da. Here we have an assumption that the peaks generated by the precursor ion should not be far away from the m/z value reported by the high-resolution MS instrument.

The maximum charge state of isotopic envelope candidates Z_{max} is the maximum between the reported precursor charge state and a user-defined maximum allowed threshold. The upper bound of Z_{max} guarantees the algorithm can be finished in limited number of cycles.

For each candidate charge state z ($1 \leq z \leq Z_{max}$), we try to find all isotopic envelopes with peak intra-spacing $\frac{1}{z}$. Therefore, the following steps are repeated Z_{max} times:

1. Set a temporary peak list $PL_{tmp} = PL$. Set an isotopic envelope candidate list $CL = \{\Phi\}$;

2. All the peaks in the peak list PL_{tmp} are scanned one by one from low m/z value to the highest one:

   (a) Suppose the first scanned peak is $P_k, k = 0, 1, ...$, and P_k is treated as the mono-isotopic peak of a new isotopic envelope E_i: $P_{i,0} = P_k$. All the peaks following $P_{i,0}$ with the step interval $\frac{j}{z}$ (j is the step number $j = 1, 2, ...$) are selected as $P_{i,0}$'s isotopic peaks. Knowing the m/z value, intensity and charge state z of $P_{i,0}$, we can compute the theoretical intensities of all isotopes of $P_{i,0}$ by using the look-up table, to help distinguish $P_{i,0}$'s real isotopes among neighbouring noise peaks. A normalized error ratio $r_{i,j}$ is computed to evaluate the intensity deviation between the j^{th} isotopic peak $P_{i,j}$ and its theoretical intensity:

   $$r_{i,j} = \frac{\text{Theoretical intensity of } P_{i,j} - \text{Observed intensity of } P_{i,j}}{\text{Theoretical intensity of } P_{i,j}} \tag{4.1}$$

   Due to the fact that there might be a few peaks appearing in the isotope m/z range, the lowest $r_{i,j}$ guarantees the selection of the best matched peak. Thus, we get an isotopic envelope candidate $E_i = \{P_{i,0}, P_{i,1}, ..., P_{i,m}\}$, where $P_{i,m}$ is the last peak that satisfies the peak intra-spacing requirement.

   (b) Remove $P_{i,0}, P_{i,1}, ..., P_{i,m}$ from the peak list PL_{tmp}, and add E_i to the isotopic envelope candidate list CL.

   (c) Then we move to the next peak in PL_{tmp}, repeat the previous two steps (a) and (b), until either we reach the end of PL_{tmp} or the m/z value of the next peak is greater than mono-isotopic peak threshold $m + 0.001m$.

3. Set $z = z + 1$, and repeat steps 1-3 until $z > Z_{max}$.

Besides the isotopic envelopes selected directly from the peak list, isotopic envelope candidates are also generated from their sub-envelopes. A sub-envelope of E_i is an envelope that contains consecutive peaks in E_i, but starts from an isotope rather than the mono-isotope. For example, $E_i^j = \{P_{i,j}, P_{i,j+1}, \ldots, P_{i,m}\}$ is j^{th} sub-envelope of E_i, and we call E_i the parent envelope of its sub-envelopes. If the m/z value of $P_{i,j}$ is greater than the right boundary of the mono-isotope threshold, the sub-envelope generation stops after generating E_i. Generating sub-envelopes guarantees that all the potential isotopic envelopes, especially the isotopic envelopes overlapped with others, are provided before the evaluation. We check the mono-isotopic m/z value of each envelopes and their sub-envelopes, only whose value within the range from $m - 0.001m$ Da to $m + 0.001m$ Da are left as final candidates.

4.1.3 Candidate Evaluation

In the last section, the approach for potential isotopic envelope candidate prediction was introduced. In this section, I will present the way to choose a candidate that best matches the observed peak distribution in the selected MS spectrum window. The selection can be done by comparing each candidate's theoretical isotopic peak distribution to the observed distribution.

As introduced earlier, given the mass value, the theoretical isotopic peak distribution of an ion can be calculated by checking the look-up table. Since the look-up table only provides a relative intensity distribution pattern, so initially we need to find a parameter a that transforms a candidate's theoretical intensity pattern to the observed one by scaling it. The value of parameter a can be calculated by using the least squares fitting approach to minimize the following objective function:

$$transformErr(E_i) = \min\{\sum_{j=1}^{|E_i|} (I_{i,j} - a \times theoryH(P_{i,j}))^2\} \tag{4.2}$$

where $|E_i|$ stands for the number of isotopic peaks in the isotopic envelope E_i. $P_{i,j}$ is the j^{th} isotope of E_i, $I_{i,j}$ is the observed peak intensity of $P_{i,j}$, and the function $theoryH(P_{i,j})$ returns the theoretical intensity of $P_{i,j}$ as the j^{th} isotope in E_i.

Since the value of a has been computed in formula 4.2, we substitute a as a_0. After scaling the intensities of E_i's theoretical pattern by the value a_0, we can compute the fitting error between the theoretical pattern and all peaks in the selected window as follows:

$$\begin{aligned} fitErr(E_i) &= \sum_{n=n_1}^{n_{|E_i|}} (I_n - a_0 \times theoryH(P_n))^2 + \sum_{m=m_1}^{m_{N-|E_i|}} (I_m)^2 \\ &= transformErr(E_i) + \sum_{m=m_1}^{m_{N-|E_i|}} (I_m)^2 \end{aligned} \tag{4.3}$$

where $P_{n_1}, P_{n_2}, ..., P_{n_{|E_i|}} \in E_i$, while $P_{m_1}, P_{m_2}, ..., P_{m_{N-|E_i|}} \notin E_i$. N represents the number of peaks in the selected window in the MS spectrum, and $N - |E_i|$ stands for the number of peaks that are not in E_i.

The candidate with the smallest fitting error (i.e., the largest similarity) compared to the selected window is considered as the precursor ion. Before accepting this precursor ion,

we need to compute the deviation between its theoretical pattern and its observed pattern to see how well it fits:

$$\| \frac{\sum_{j=1}^{|E_i|} (a_0 \times theoryH(P_{i,j}))^2 - \sum_{j=1}^{|E_i|} (I_{i,j})^2}{\sum_{j=1}^{|E_i|} (a_0 \times theoryH(P_{i,j}))^2} \| \leq I_{threshold} \tag{4.4}$$

where $|E_i|$ represents the number of the isotopic peaks in E_i. $P_{i,j}$ is the j^{th} isotope of E_i, and $I_{i,j}$ is the observed peak intensity of $P_{i,j}$.

If the deviation of the candidate satisfies the inequality 4.4 , then the selected precursor ion is accepted; otherwise, the algorithm returns null and will not correct the reported precursor mass and charge value. $I_{threshold}$ is a threshold value, which is given based on expertise.

An example is given in Figure 4.4 to describe each step in determining the mono-isotopic m/z value and charge state of the precursor ion.

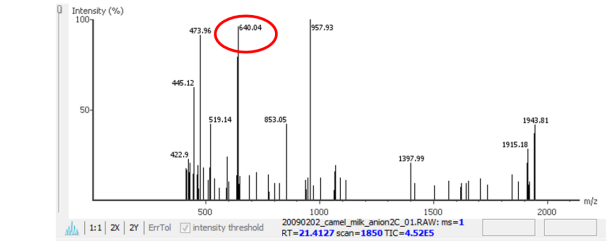
4.2 MS/MS Spectrum Deconvolution Problem

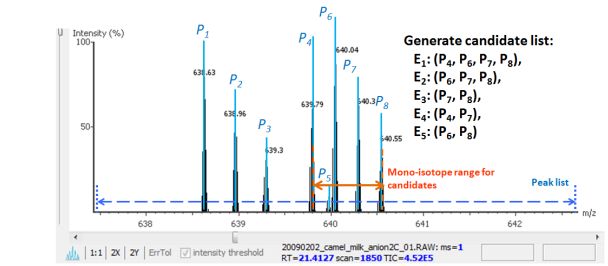
4.2.1 MS/MS Spectra Deconvolution Problem: An Application of Charge and Mass Correction Algorithm

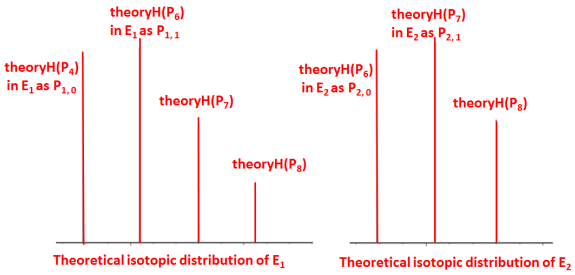
If the precursor ion charge and mass correction algorithm is applied to the entire spectrum, rather than a selected window of the spectrum, it could be used to solve a more challenging problem: the MS/MS spectrum pre-processing problem. This problem requires a more complicated algorithm that converts a spectrum full of multiple charged ions as well as isotopes into a spectrum that contains only single charged mono-isotope signals.

A widely used strategy, called "*Deconvolution*", has been proposed to interpret the MS spectra or MS/MS spectra. It contains two major steps:

- Move a window along the m/z mass value and identify all the isotopic envelopes of unknown ion in each window.
- For each of the identified isotopic envelopes, do de-isotope and multiple charge deconvolution.







$$fitErr\ (E_i) = \min\{\sum_{n=n_1}^{n_{|E_i|}} (I_n - a \times theoryH\ (P_n))^2 + \sum_{m=m_1}^{m_{N-|E_i|}} (I_m)^2$$

$$\left\| \frac{\sum_{j=1}^{|E_i|} (a_0 \times theoryH\ (P_{i,j}))^2 - \sum_{j=1}^{|E_i|} (I_{i,j})^2}{\sum_{j=1}^{|E_i|} (a_0 \times theoryH\ (P_{i,j}))^2} \right\| \leq I_{threshold}$$

Candidate with minimum fitting error:
E_1= {P_4, P_6, P_7, P_8}
Precursor ion:
m/z value: mz(P_4)=*639.74 Da*
Charge state: *z = 4*

Collect the peaks in the parent spectrum within m/z range:
[*m - 5, m + 5*] Da

Produce a candidate list *CL* which contains isotopic envelopes with
mono-isotopic m/z within range:
[*m - 0.001m, m + 0.001m*] Da

Generate theoretical isotopic distribution of each candidate in *CL*

Compute the fitting error between every candidate and the MS spectrum window

Evaluate the quality of the selected precursor ion candidate

Output the precursor ion if it is accepted

Figure 4.4: An example: steps to determine the mono-isotopic mass and charge state.

Obviously, the key part is the first step: accurately identify all isotopic envelopes for each unknown ions. In a spectrum, isotopic envelopes may overlap each other, and the algorithm should detect as many envelopes as possible. So far, several algorithms have been proposed to automatically deal with spectra data preprocessing. THRASH [46] and some related algorithms [47] [48] are designed for automatically interpreting top-down high resolution MS spectra for structural characterization analysis of large protein molecules. MS spectra usually contain a variety of highly charged peptides and a huge amount of noise. However, these two sorts of signals are not frequently observed in MS/MS spectra, instead, the mono-isotope determination and isotopes overlapping have become the major problems in MS/MS spectra. Therefore, it is not suitable to directly apply these deconvolution algorithms to the MS/MS spectra. Sean McIlwain and et al. [49] introduced a method, called Isotope Distribution Mapper (IDM), which used a probabilistic classifier to identify isotopic distribution on smoothed data. IDM can be trained to make it more robust that can be used on data from a wider array of experimental conditions, but so far it cannot handle the isotope overlapping problem.

In this section, I will introduce a deconvolution algorithm for MS/MS spectra based on our charge and mass correction algorithm.

4.2.2 Method Overview

Figure 4.5 illustrates the overall procedure of our deconvolution algorithm. Given a MS/MS spectrum S, our algorithm applies the following four steps to do the pre-processing:

1. Remove noise and do peak centroiding.
2. Divide the entire peak list into several consecutive groups: $S = \{G_1, G_2, G_3, ..., G_m\}$. A group G_i is a set of neighbouring isotopic envelopes. These envelopes may be overlapped, however, every group does not contain any two consecutive peaks with intra distance greater than 1.0086 Da, which is the maximum m/z difference of two adjacent isotopes.
3. Set the isotopic envelope list $EL = \{\Phi\}$. For each group G_i in S:
   (a) Generate a candidate list containing all the possible isotopic envelopes.
   (b) Compute a least squares fitting score for every combination of two candidates, and select the combination E_i and E_j with the least fitting error.

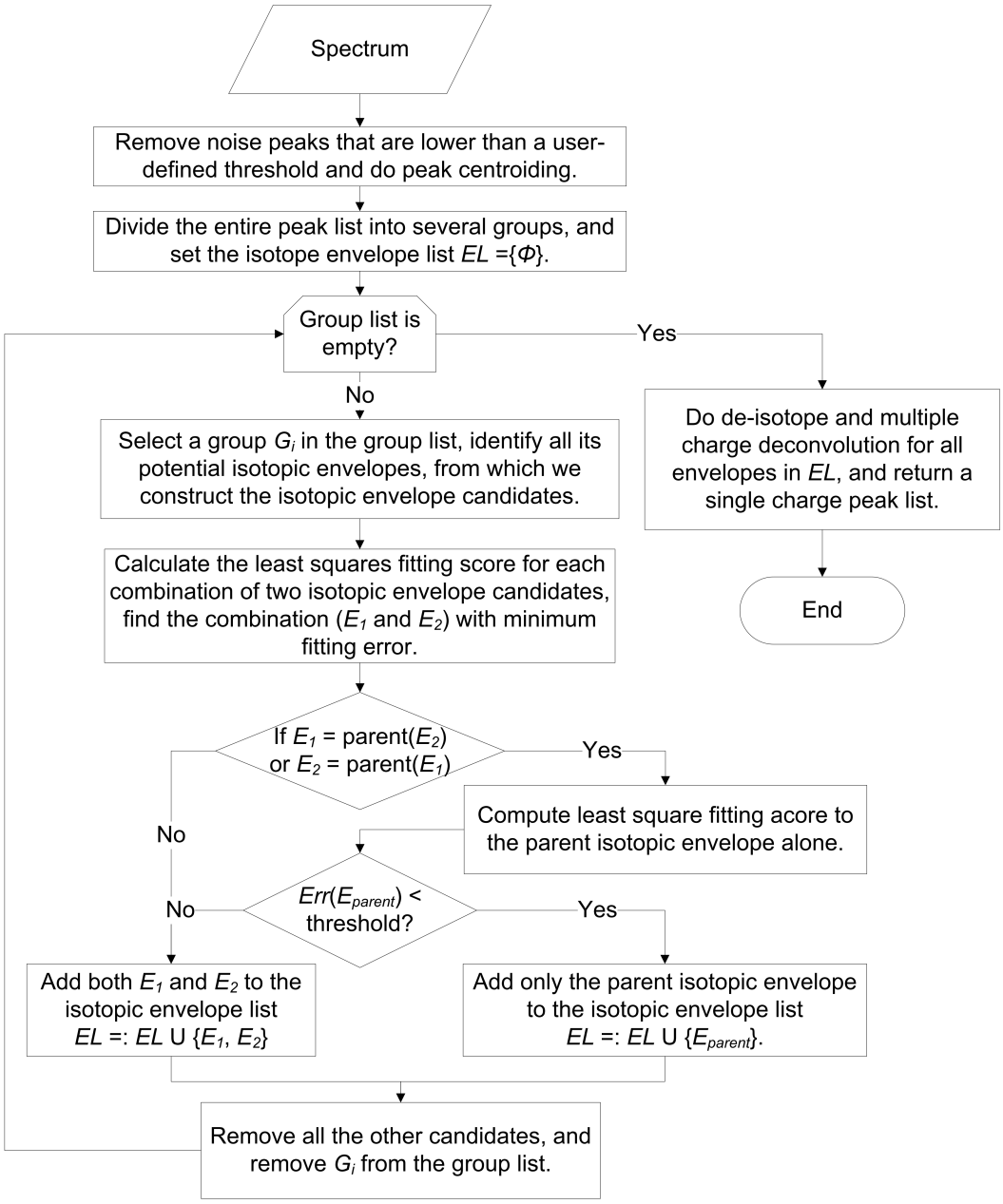


Figure 4.5: The frame work of the proposed deconvolution algorithm.

(c) If one of E_i or E_j is a subset of the other, we need to check whether E_i and E_j are truly overlapped isotopic envelopes. If the parent envelope alone can match the observed peaks well enough, then only the parent envelope is added to EL. Otherwise, EL is set to $EL \cup \{E_i, E_j\}$.

(d) Remove all the other candidates, and remove G_i from S. Move to the next group, repeats steps (a) to (d).

4. De-isotope and do multiple-charge deconvolution for all candidates in EL and return a mono isotopic peak list.

The major difference between the precursor correction algorithm and the deconvolution algorithm is how we select candidates in a group. In the following subsections more details are given for each step.

4.2.3 Isotopic Envelopes Selection

The maximum charge state Z_{max} of each group can be determined as follows:

$$Z_{max} \leq \frac{\text{precursor mass value of MS/MS}}{\text{m/z of the first peak in } G_i} \tag{4.5}$$

For each candidate charge state z ($1 \leq z \leq Z_{max}$), the same method described in Section 4.1.2 is used to select all the possible candidates.

4.2.4 Fitting Score for Choosing the Best Candidates

So far, all of the potential isotopic envelope candidates with different charge states have been prepared for further filtering. The general idea of our algorithm is: select any two of the candidates to fit the peak intensity distribution of the given group G_i. We apply a least squares fitting method to each combination, and the one with the minimum error is selected.

There are two reasons why we consider only the best two isotopic envelopes in one group.

First, different from MS spectra, MS/MS spectra do not contain dense peak signals. In most groups observed, only one major isotopic envelope can be found. Other small

surrounding peaks are more like noise that does not have obvious shape features, such as width, intensity and intra peak space. Even for human experts, it is usually hard to group them together. Choosing the top one or two isotopic envelopes in a group enables us to enhance the real signals and weaken the noise to the largest extent, as a consequence benefitting both *de novo* and database search approaches.

Secondly, choosing two isotopic envelopes can effectively solve the overlapping problem. We observe a number of high-intensity isotopic envelopes whose abundances of peaks are quite different from their theoretical abundance distribution determined by the mass of the mono-isotope peak. After carefully analyzing, we believe that these strange abundances of peaks are caused by the overlapping of two ions with one Dalton mass offset and the same charge state. This is possible. For example, a portion of ion I loses a water while another portion of I loses an ammonia during the fragmentation process. The produced two variant ions : $I - H_2O$ and $I - NH_3$ together will generate an isotopic envelope with the same peaks compared to the one generated by ion $[I - H_2O]$ alone, but with quite different peak abundance distribution. Moreover, the probability of the existence of two such similar-mass-value ions is higher than the probability of three or more ions whose existences highly depend on the component of the parent molecule and may not be frequently observed.

Let $G_i = \{P_1, P_2, \ldots, P_k\}$, where k is the number of peaks in G_i. Let $I^{obs}_{G_i}$ be the observed intensity vector of G_i: $I^{obs}_{G_i} = \{I_1, I_2, \ldots, I_k\}$, where I_i is the intensity of P_i. Given two isotopic envelopes E_1 and E_2, we also use intensity vectors to represent their theoretical distributions: $I^{thy}_{E_1} = \{x_1, x_2, \ldots, x_k\}$, $I^{thy}_{E_2} = \{y_1, y_2, \ldots, y_k\}$, where x_i and y_i are the theoretical isotopic peak intensities of E_1 and E_2, respectively. If a peak $P_c \notin E_1$, let $x_c = 0$. Similarly, if $P_c \notin E_2$, let $y_c = 0$. Now we need to find two parameters a and b to satisfy:

$$\begin{cases} a \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_k \end{pmatrix} + b \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_k \end{pmatrix} = \begin{pmatrix} I_1 \\ I_2 \\ \ldots \\ I_k \end{pmatrix} \\ a \geq 0 \\ b \geq 0 \end{cases} \tag{4.6}$$

Such an equation may not have a solution when $k > 2$. In such case, we try to minimize

the following objective function:

$$Err(E_1, E_2) = min \sum_{c=1}^{k} (ax_c + by_c - I_c)^2 \tag{4.7}$$

This can be solved by a standard least squares algorithm. The only change here is that if a (or b) is negative, then we set $a = 0$ (or $b = 0$) and apply the least squares algorithm again to calculate b (or a) only. It indicates that one of the envelope candidates is adequate to explain the component in G_i, instead of the combination of the two.

After calculating all the combinations, the one with minimal fitting error is obtained. Then we check the two isotopic envelopes E_1 and E_2, to see whether there exists an inclusion relationship between them: if $E_2 \subset E_1$, then we let E_1 be the parent envelope of E_2. If such a relationship exists, re-checking is required to see whether both E_1 and E_2 are assigned a reasonable amount of peak intensities in G_i:

$$|\frac{Err(E_{parent}) - Err(E_1, E_2)}{Err(E_{parent})}| \geq \theta_{threshold} \tag{4.8}$$

where $Err(E_{parent})$ is the least squares fitting error for the parent isotopic envelope. If the above inequality is not satisfied, then only the parent isotopic envelope will be kept in G_i; otherwise, both of them will be retained. After the above steps, all the other candidates are removed from group G_i.

If a given group G_i contains only two peaks, sometimes it may be difficult to decide their charge states. In this case, we would check whether the isotopic envelope(s) derived from the same ion but with different charge states exist. If there is such an envelope, and these two peaks in G_i follow their theoretical isotopic distribution, we consider them as one isotopic envelope; otherwise, consider them as coming from two single ions.

4.2.5 Multiple Charge Deconvolution

De-isotopes and remove multiply charged ions are done on every isotopic envelope E_i in EL as follows:

- **De-isotopes:** Accumulate the intensities of isotopes as the total intensity of E_i.

$$intensity(E_i) = \sum_{j=1}^{|E_i|} I_{i,j}$$

where $|E_i|$ stands for the number of the isotopic peaks in the isotopic envelope E_i. $I_{i,j}$ is the observed peak intensity of the j^{th} isotope of E_i.

- **Remove multiply charged ions:** Compute the m/z value of the singly charged ion from which E_i was derived. If other ion(s) has quite similar m/z value, i.e. within a given mass error tolerance, we treat them as the same ion. We sum up their intensities and only one peak signal at that m/z site is left to represent this ion.

4.3 Experiments and Results

4.3.1 Experiments for Precursor Mono-Isotopic Mass and Charge Correction Algorithm

The precursor mono-isotopic mass and charge correction algorithm was tested with two standard protein mixtures. The standard mixtures were reduced and alkylated by iodoacetamide, then digested by trypsin overnight. The peptide mixture (2 μL injected) was separated via Surveyor^{TM} LC equipped with MicroAS^{TM} autosampler (Thermo Fisher Scientific) using a reversed phase analytical column (75 μm inner diameter, 10 cm length, 3m particle size, both Nanoseparations, NL), at a flow rate of 250 nL/min. A gradient of 5 - 30% acetonitrile in 90 minutes was used. Both data sets were obtained from Thermo LTQ-Orbitrap XL:

- The first data set contains 6317 MS spectra and 711 MS/MS spectra. The precursor mono-isotopic m/z values of 100 spectra were manually annotated by a human expert (the annotator). The annotation showed that 68 out of the 100 precursor m/z reported by the instrument were not mono-isotopic peak.
- The second data set contains 32 MS spectra and 96 MS/MS spectra. All 96 precursor masses reported by the instrument were not mono-isotopic peak.

4.3.2 Results

The experimental results were examined by a human expert and shown in Figure 4.6. Figures (a) and (b) illustrate the precursor ion mass shift compared with expert's annotation before and after the correction algorithm on the first data set. Out of the 100 annotated spectra, 68 are inconsistent with expert's annotation before the correction. After applying

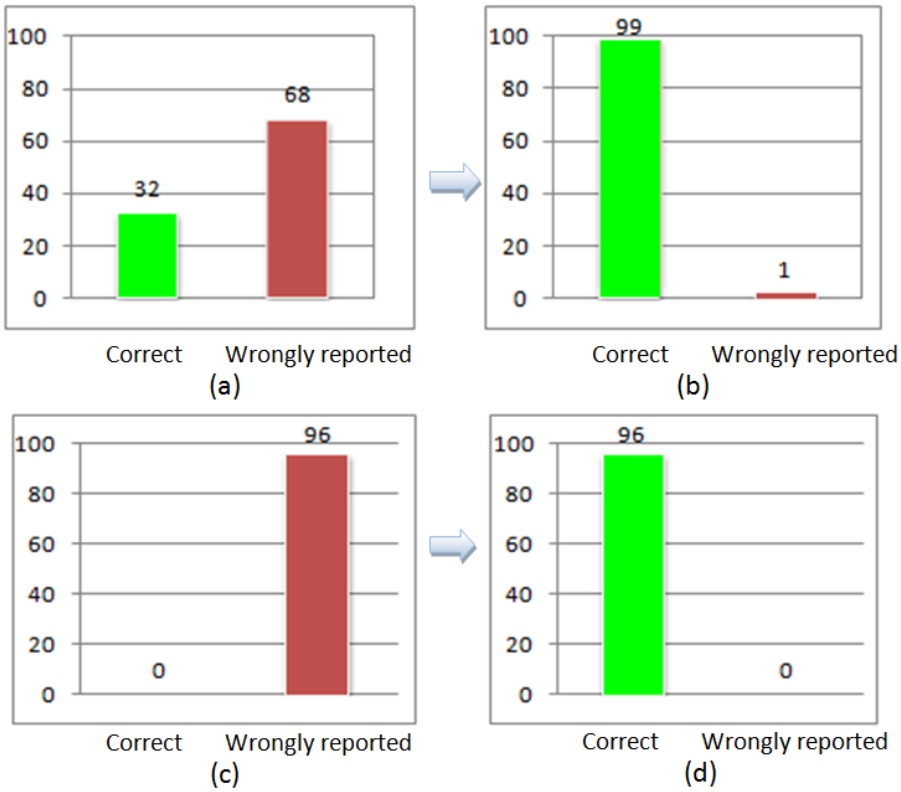


Figure 4.6: Experiment on data set A: precursor mono-isotopic peak correction compared with human expert's annotation: (a) before correction vs. (b) after correction; experiment on data set B: (c) before correction vs. (d) after correction.

our algorithm, 67 out of 68 inconsistent spectra have been corrected and the remaining 32 spectra stayed unchanged, so the result was consistent with the human annotation for 99 (out of 100) in total. For the one inconsistent spectrum, the annotator agreed that the software's calculation was more trustworthy after re-examining the data.

For the second testing data set (as shown in Figure 4.6(c) and (d)), 96 (out of 96) spectra were corrected and were consistent with the human expert's annotation.

4.3.3 Experiments for MS/MS Spectra Deconvolution Algorithm Compared with Two Other Software Tools

Our MS/MS spectra deconvolution algorithm was tested on a data set downloaded from an online proteomics database [43]. The data was derived from a mixture of yeast proteins. It was digested with trypsin and measured with CID fragmentation modes of an LTQ Orbitrap XL (Thermo Fisher Scientific^{TM}, Bremen, Germany). It contains 2172 MS and 694 MS/MS (m/z 100-2000 Da). The high-confidence peptide sequences identified from the yeast data set were obtained from a consensus result by two database search tools: PEAKS 5.1 and Mascot. In total, we selected 124 high-confident PSMs which satisfied the following three conditions:

1. Both database search tools reported the same peptides;
2. The PEAKS confidence score was not less than 0.9;
3. The Mascot confidence score was not less than 0.6.

These 124 PSMs were used as a control set to test the performance of our deconvolution software and two other tools: the "Data Refine" function of PEAKS 5.1 and the Isotope Distribution Mapper (IDM). The "Data Refine" function can export a mono-isotopic peak list as our algorithm does. IDM does not provide a mono-isotopic peak list, instead, it gives an isotopic distribution map which includes all the predicted isotopic envelopes in each spectrum. Therefore, the multiple charge deconvolution approach mentioned in the section 4.2.5 was used to obtain a mono-isotopic peak list from the identified isotopic envelopes of IDM; all of the remaining peaks were treated as single charged peaks, and a mono-isotopic peak list was generated from them; two mono-isotopic peak lists were combined as the deconvoluted result of IDM.

Thus, for any spectrum in the control set, we compared the three deconvoluted spectra generated by these software tools in terms of the ability to identify peptides.

Experimental Procedure

Two experiments were designed:

- **Compare the amount of correctly identified fragmentation ions:** In CID MS/MS spectra, the most frequently observed ions are: y-ion, b-ion, y-ion$-H_2O$,

b-ion$-H_2O$, y-ion$-NH_3$, b-ion$-NH_3$ and a-ion. Given a spectrum and its real sequence, a theoretical spectrum containing only these seven kinds of ions can be produced. To compare three tools in a fair way, we selected the highest top K peaks of each deconvoluted spectra, and computed the number of matches between them and the theoretical spectrum.

- **Compare the number of correctly identified peptides by using a *de novo* sequencing approach:** PEAKS 5.1 was used to do *de novo* sequencing on the deconvoluted spectra processed by three algorithms, respectively. The precursor mass error tolerance and fragment error tolerance were set to 0.1 Da and 0.5 Da, respectively.

4.3.4 Results

Figure 4.7 illustrates the number of matched ions found by increasing the number of selected top-intensity peaks K in each spectrum. The x-coordinate is the number of the selected peaks K, and the y-coordinate is the number of matches between these K peaks and the seven kinds of theoretical ions. The solid line represented the result of our algorithm, the thick dashed line and the thin dashed line represents the result of the Data Refine function and the IDM software, respectively. We observe that our algorithm can match more theoretical ions than the other software tools, and when K was among the top 60 to 100 peaks per spectrum, the difference was the most obvious. After K exceeded 120, the differences in matched ions became less obvious, since more and more noise peaks were included in the highest top K peaks. This figure shows that our algorithm can effectively enhance the intensity of real peaks and distinguish them despite the noise peaks.

Table 4.2: The comparisons of correctly and partial correctly identified peptides by using a *de novo* sequencing approach on three sets of the pre-processed spectra.

	IDM	PEAKS Data Refine	Our Algorithm
Correct peptides	24.39%	36.59%	36.59%
Peptides with ≤ 1 incorrect residue	25.2%	37.40%	38.21%
Peptides with ≤ 2 incorrect residues	39.02%	50.41%	51.22%
Peptides with ≤ 3 incorrect residues	42.28%	53.66%	54.47%
Total correct residues	56.21%	64.32%	65.40%

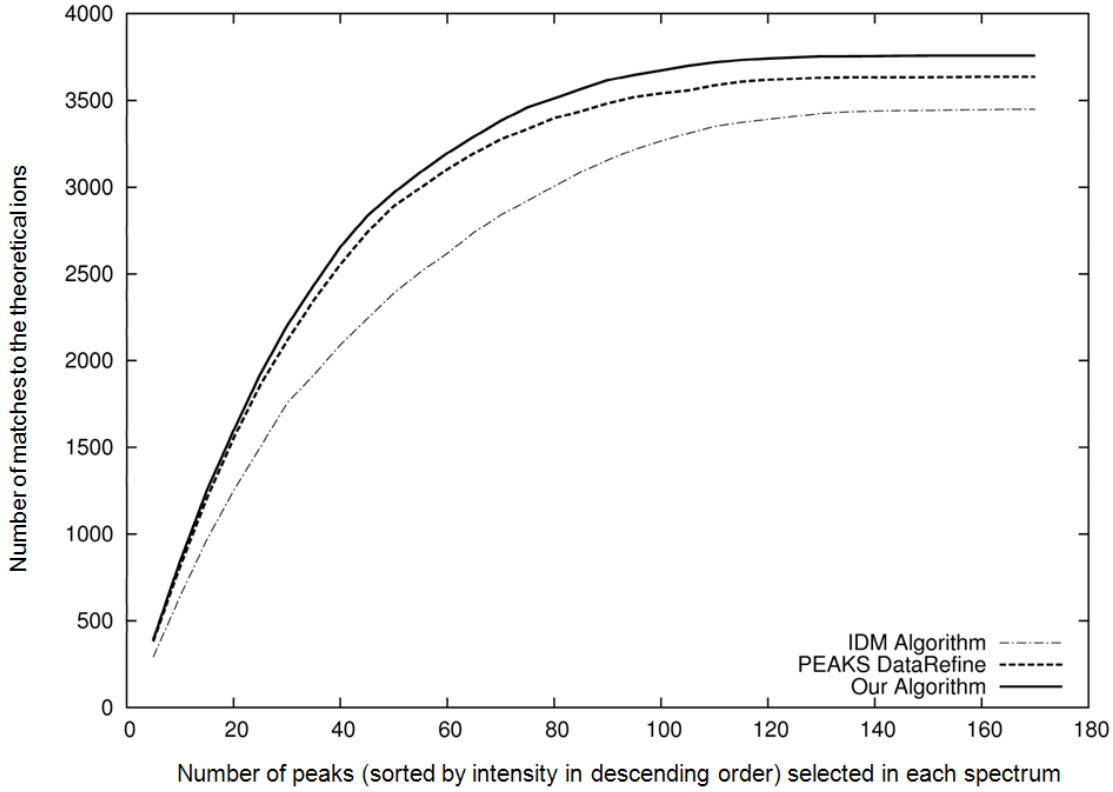


Figure 4.7: The number of matched ions found by increasing the number of selected peaks in each spectrum.

Table 4.2 illustrated the accuracy of *de novo* sequencing results after utilizing the deconvoluted spectra from three software tools. The first four rows are the comparison between the percentage of the peptides with at most 0, 1, 2, and 3 incorrect residues in the *de novo* sequencing results. The last row shows the difference among the percentage of the correctly determined residues. Using the deconvoluted spectra generated by our algorithm and the Data Refine function, the *de novo* sequencing tool can identify the same number of entirely correct peptides, which is 36.59% of the peptide identification results. However, when we allow 1, 2, or at most 3 incorrect residues on a peptide, better *de novo* sequencing results can be seen on the spectra generated by our algorithm when compared with the spectra by the other two approaches. Additionally, our algorithm obtains the best result on the total correct residues. Since PEAKS *de novo* sequencing algorithm was trained based on its' own Data Refine tool, more improvement is possible if its training is based on data pre-processed with our algorithm.

Figure 4.8 gives an example that our deconvolution algorithm is able to handle the overlapped isotopic envelopes. Figure 4.8 (a) shows the observed peak distribution within a mass range from 514 Da to 517 Da. After centroiding, each peak converts to a single signal shown in (b). The intensity ratio of the second peak to the first peak in (b) shows that overlap exists in this region. Two isotopic envelopes E_1 and E_2 were identified by our algorithm, and (c) and (d) illustrates the intensities assigned for E_1 and E_2, respectively. From (f) we can see clearly that the overlapped isotope distribution identified by our algorithm is quite close to the observed distribution. Moreover, these two envelopes are also found to be matching two theoretical ions: $\{y_9 - NH_3\}$-ion and $\{y_9 - H_2O\}$-ion, which further proves the correctness of identifications of E_1 and E_2.

4.4 Discussion

Accurately interpreting ions (i.e., the isotopic envelopes) is always an important task in the MS-based Proteomics. In this chapter, we introduced two approaches that apply isotopic envelope selection technology: the precursor mono-isotopic mass and charge correction algorithm, and the MS/MS deconvolution algorithm. The first algorithm examines the corresponding window of the parent MS spectra and selects the most probable isotopic envelope of the precursor ion. The second algorithm searches the entire MS/MS spectrum, detects all the probable isotopic envelopes, and produces a single charged mono-isotopic list as the deconvoluted spectrum. Both algorithms aim to reduce the error and the complexity of the spectra data, and further improve the results of peptide identification. Experiments show that our precursor mono-isotopic mass and charge correction algorithm can achieve

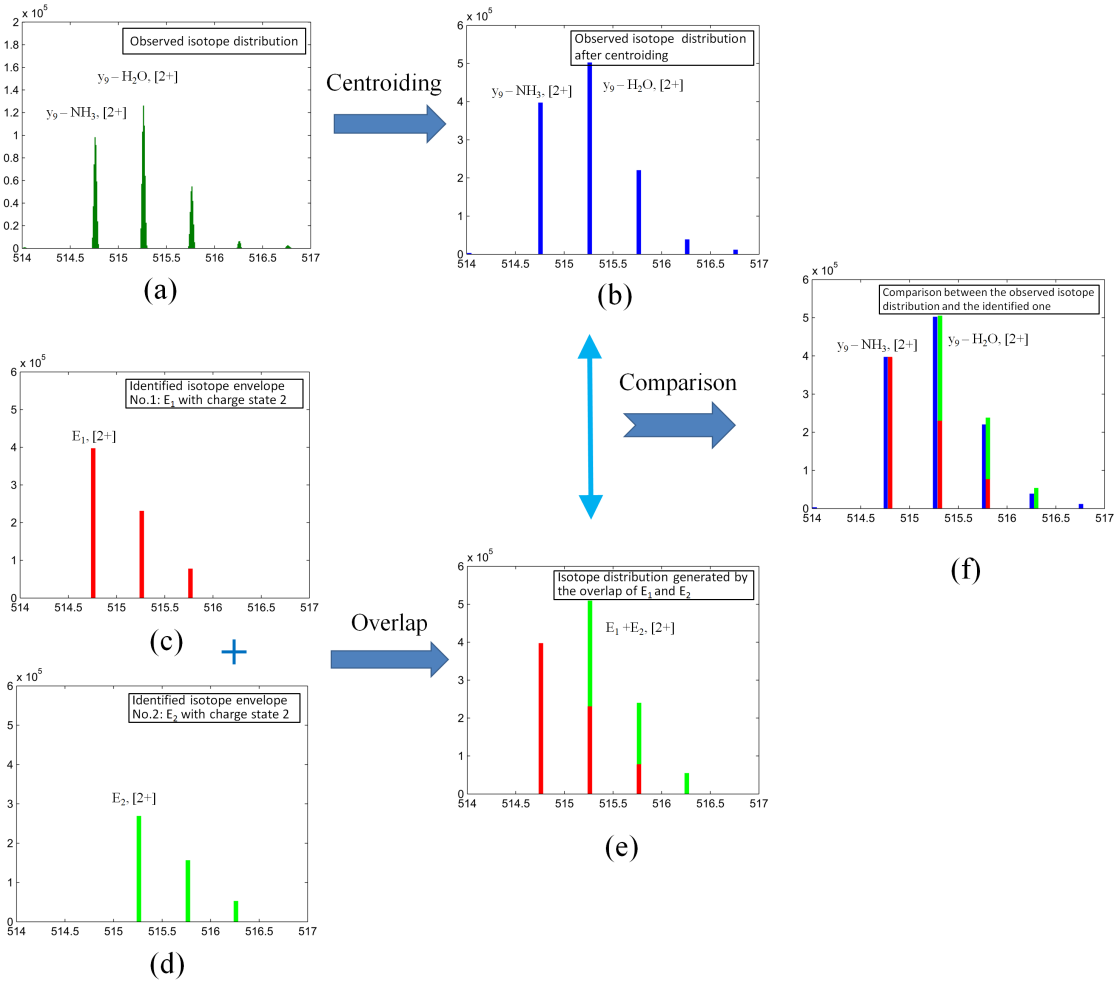


Figure 4.8: An example to show that our algorithm can handle the overlapping problem. (a) The observed peak distributions within an m/z region of a tandem MS spectrum. (b) Peaks intensity distribution after centroiding. From this distribution, our algorithm identified two overlapped isotopic envelopes (E_1 and E_2) with charge state 2. E_1 started from the peak with m/z 514.75 Da (c), and E_2 started from the highest peak with m/z 515.25 Da (d). (c) and (d) also illustrates their assigned intensities, respectively. (e) The isotope distribution generated by the overlapping E_1 and E_2. (f) A comparison to show that the identified peak distribution by our algorithm is very close to the observed peak distribution. E_1 and E_2 were found matching theoretical ions: $\{y_9 - NH_3\}$-ion and $\{y_9 - H_2O\}$-ion, respectively.

almost 100 percent accuracy on two annotated data sets, and this software has been implanted into PEAKS 5.2. Our deconvolution algorithm also performed better than two other software tools, IDM and Data Refine function in PEAKS 5.1. As commercial software, PEAKS 5.1 also got good results during the comparison, while our algorithm still gained a slight advantage.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The specific contributions and conclusions of this study can be summarized as follows:

- PeaksPTM, a software tool for identifying peptide sequences with unspecified modifications is proposed in this thesis. PeaksPTM adopts a two-pass strategy and is able to invoke all the known post-translation modifications characterized in Unimod database. To increase the confidence of the identification of a modified peptide, the new scoring function utilizes two features: peptide pair and PTM rareness, to improve the unrestricted PTM search. A modified target-decoy database strategy is used in PeaksPTM to estimate the false discovery rate. Two experiments are provided to show that PeaksPTM significantly outperforms four other major unrestrictive search engines. A consensus experiment confirms the evidence in the literature that inefficiencies in modified peptide identification is one of the major factors for the low characterization rate of the MS/MS spectra in a data set and the low identification rate of the modified peptides. The PeaksPTM software is freely accessible at http://www-novo.cs.uwaterloo.ca:8080/PeaksPTM/.

- We propose a precursor mono-isotopic mass and charge determination approach to further improve the accuracy of the peptide identification process. The key idea is to predict a theoretical isotopic pattern for each potential candidate in a window of the parent spectrum and select the best matched one. Experimental results show that our approach can achieve almost 100 percent accuracy on two annotated data sets.

We also apply this approach to the MS/MS preprocessing problem, and propose a deconvolution algorithm that can handle the overlapping of isotopic envelopes. Experimental results show that our deconvolution algorithm can effectively detect the potential ions in the MS/MS spectra, and perform better than two other existing software tools.

5.2 Future Work

The objective of this thesis is to study the strategies to improve peptide identification in MS-based proteomics, especially the identification of modified peptides. One important part is to study the ubiquitous incorporation of PTMs on the peptides, and there remain many problems in our research:

- First, in the current version of PeaksPTM, although we only use the most frequently observed 10 to 20 modifications for the identification of peptides with multiple PTMs, the generation and scoring for a peptide candidate with multiple PTMs are still time consuming. An immediate improvement of PeaksPTM is to design a new algorithm for the peptide candidate generation, with the goal of producing more reasonable candidates rather than trying all the possible PTM combinations.
- Second, our current work focuses on PTM identification, while in reality, the PTM site localization is at least as important as PTM identification. Evaluation of the confidence of each identified modification, and detection of an inaccurately assigned modification site, still need to be intensively studied in the future.

References

[1] B. Ma. Challenges in computational analysis of mass spectrometry data for proteomics. *Journal of Computer Science and Technology*, 25(1):107–23, 2010. x, 8

[2] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS: powerful software for MS/MS peptide de novo sequencing. *Rapid Commun. in Mass Spectrom*, 20:2337–2342, 2003. 1, 9, 10, 22

[3] J. U. Baenziger. A major step on the road to understanding a unique posttranslational modification and its role in a genetic disease. *Cell*, 113:421–422, 2003. 1, 9, 18

[4] M. Mann and O. Jensen. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, 21(3):255–261, 2003. 1, 2, 9, 18, 24

[5] E. S. Witze, W. M. Old, K. A. Resing, and N. G. Ahn. Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, 4(10), 2007. 1, 2, 9, 18

[6] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999. 1, 9, 18

[7] A. I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. of Proteomics*, 73(11):2092–2123, 2010. 1, 4

[8] A. I. Nesvizhskii, F. F. Roos, J. Grossmann, M. Vogelzang, J. S. Eddes, and W. Gruissem et al. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell Proteomics*, 5:652–670, 2006. 1

[9] R. J. Chalkley, P. R. Baker, L. Huang, K. C. Hansen, N. P. Allen, and M. Rexach et al. Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer –ii. new developments in protein prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol. Cell Proteomics*, 4:1194–1204, 2005. 1

[10] M. L. Nielsen, M. M. Savitski, and R. A. Zubarev. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol. Cell Proteomics*, 5:2384–2391, 2006. 1

[11] S. Tanner, H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77:4626–4639, 2005. 1, 11, 26, 28, 36

[12] J. R. Yates III., J. K. Eng, A. L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, 67(8), 1995. 1

[13] D. M. Creasy and J. S. Cottrell. Unimod: Protein modifications for mass spectrometry. *Proteomics*, 4(6):1534–1536, 2004. 2

[14] ABRF Delta Mass database. `http://www.distributed-generation.com`. Visited at Nov. 2010. 2

[15] F. Wold. In vivo chemical modification of proteins. *Ann. Rev. Biochem.*, 50:783–814, 1981. 2

[16] M. Duncan, R. Aebersold, and R. Caprioli. The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.*, 28(7):659–664, 2010. 2, 10, 36

[17] J. U. Baenziger. A major step on the road to understanding a unique post translational modification and its role in a genetic disease. *Cell*, 113:421–422, 2003. 2, 10

[18] X. Han, L. He, L. Xin, B. Shan, and B. Ma. PeaksPTM: Mass spectrometry based identification of peptides with unspecified modifications. *J. Proteome Res.*, 10(7):2930–2936, 2011. 3

[19] W. P. Blackstock and M. P. Weir. Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*, 17(3):121–127, 1999. 4

[20] N. L. Anderson and N. G. Anderson. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19(11):1853–61, 1998. 4

[21] M. R. Wilkins et.al. From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and arnino acid analysis. *Nature Biotechnology*, 14(1):61–65, 1996. 4

[22] V. C. Wasinger et. al. Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis*, 16(1):1090–1094, 1995. 4

[23] J. A. Taylor and R. S. Johnson. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 11(9):1067–1075, 1997. 10

[24] V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry: a graph-theoretical approach. *J. Comp. Biol.*, 6:327–342, 1999. 10

[25] A. Frank and P. Pevzner. Pepnovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, 77(4):964–973, 2005. 10

[26] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, 66(24):4390–4399, 1994. 11

[27] D. L. Tabb, A. Saraf, and J. R. Yates III. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.*, 75:6415–6421, 2003. 11

[28] B. C. Searle, S. Dasari, M. Turner, A. P. Reddy, D. Choi, P. A. Wilmarth, A. L. McCormack, L. L. David, and S. R. Nagalla. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal. Chem.*, 76:2220–2230, 2004. 11

[29] Y. Han, B. Ma, and K. Zhang. SPIDER: software for protein identification from sequence tags containing de novo sequencing error. *J. Bioinformatics and Comput. Biol.*, 3(3):697–716, 2005. 11

[30] S. Kim, S. Na, J. W. Sim, H. Park, J. Jeong, H. Kim, Y. Seo, J. Seo, K. J. Lee, and E. Paek. MOD^i : A powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra. *Nucleic Acids Res.*, 34:258–263, 2006. 11, 28

[31] M. Bern, Y. Cai, and D. Goldberg. Lookup peaks: A hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.*, 79:1393–1400, 2007. 11

[32] I. V. Shilov, S. L. Seymour, A. A. Patel, A. Loboda, W. H. Tang, S. P. Keating, C. L. Hunter, L. M. Nuwaysir, and D. A. Schaeffer. The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics*, 6:1638–1655, 2007. 11, 28

[33] D. M Creasy and J. S. Cottrell. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics.*, 2:1426–1434, 2002. 11, 28

[34] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300, 1995. 13

[35] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increase confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4:207–214, 2007. 13

[36] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, 7(01):29–34, 2008. 13, 14, 30

[37] K. A. Reidegeld, M. Eisenacher, M. Kohl, D. Chamrad, G. Körting, M. Blüggel, H. E. Meyer, and C. Stephan. An easy-to-use Decoy Database Builder software tool, implementing different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications. *Proteomics.*, 8(6):1129–1137, 2008. 13, 14

[38] A. L. Rockwood, S. L. Van Orden, and R. D. Smith. Rapid calculation of isotope distributions. *Anal Chem*, 67:2699–2704, 1995. 16

[39] M. Bern, B. S. Phinney, and D. Goldberg. Reanalysis of tyrannosaurus rex mass spectra. *J. Proteome Res*, 8:4328–4332, 2009. 20, 27

[40] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. A. Pevzner. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.*, 23:1562–1567, 2005. 22

[41] M. M. Savitski, M. L. Nielsen, and R. A. Zubarev. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of

modifications, and fingerprinting complex protein mixtures. *Mol. Cell. Proteomics*, 5:935–947, 2006. 22

[42] D. J. Graves, B. L. Martin, and J. H. Wang. *Co- and post-translational modification of proteins: chemical principles and biological effects.* Oxford University Press, 1994. 24

[43] Proteome Commons database. `https://proteomecommons.org/`. 29, 57

[44] A. Keller, A. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, 74:5383–5392, 2002. 30

[45] M. W. Senko, S. C. Beu, and F. W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Sot. Mass Spectrom.*, 6:229–33, 1995. 45

[46] D. M. Horn, R. A. Zubarev, and F. W. McLafferty. Automated reduction and interpretation of high resulution electrospray mass spectrua of large molecules. *J. Am. Soc. Mass Spectrom*, 11:320–332, 2000. 45, 50

[47] C. Li, K. S. Siu, and Y. He. Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. *Anal. Chem.*, 78:5006–5018, 2006. 50

[48] P. Kaur and P. B. OĆonnor. Algorithms for automatic interpretation of high resolution mass spectra. *J Am Soc Mass Spectrom.*, 17:459–68, 2006. 50

[49] S. McIlwain, D. Page, E. L. Huttlin, and M. R. Sussman. Using dynamic programming to create isotopic distribution maps from mass spectra. *Bioinformatics*, 23:328–336, 2007. 50