

Algorithms for Characterizing Peptides and Glycopeptides with Mass Spectrometry

by

Lin He

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2013

© Lin He 2013

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.

Lin He

Abstract

The emergence of tandem mass spectrometry (MS/MS) technology has significantly accelerated protein identification and quantification in proteomics. It enables high-throughput analysis of proteins and their quantities in a complex protein mixture. A mass spectrometer can easily and rapidly generate large volumes of mass spectral data for a biological sample. This bulk of data makes manual interpretation impossible and has also brought numerous challenges in automated data analysis. Algorithmic solutions have been proposed and provide indispensable analytical support in current proteomic experiments. However, new algorithms are still needed to either improve result accuracy or provide additional data analysis capabilities for both protein identification and quantification.

Accurate identification of proteins in a sample is the preliminary requirement of a proteomic study. In many cases, a mass spectrum cannot provide complete information to identify the peptide without ambiguity because of the inefficiency of the peptide fragmentation technique and the prevalent existence of noise. We propose ADEPTS to this problem using the complementary information provided in different types of mass spectra. Meanwhile, the occurrence of posttranslational modifications (PTMs) on proteins is another major issue that prevents the interpretation of a large portion of spectra. Using current software tools, users have to specify possible PTMs in advance. However, the number of possible PTMs has to be limited since specifying more PTMs to the software leads to a longer running time and lower result accuracy. Thus, we develop DeNovoPTM and PeaksPTM to provide efficient and accurate solutions.

Glycosylation is one of the most frequently observed PTMs in proteomics. It plays important roles in many disease processes and thus has attracted growing research interest. However, lack of algorithms that can identify intact glycopeptides has become the major obstacle that hinders glycoprotein studies. We propose a novel algorithm, GlycoMaster DB, to fulfil this urgent requirement.

Additional research is presented on protein quantification, which studies the changes of protein quantity by comparing two or more mass spectral datasets. A crucial problem in the quantification is to correct the retention time distortions between different datasets. Heuristic solutions from previous research have been used in practice but none of them has yet claimed a clear optimization goal. To address this issue, we propose a combinatorial model and practical algorithms for this problem.

Acknowledgements

It has been six years since I came to Canada for this doctorate. I could have not reached it without the support and advice from my supervisor, Dr. Bin Ma. It was Bin who introduced me to this interesting interdisciplinary research area and kept fostering my development as a researcher patiently. I would like to express my heartfelt appreciation for his guidance and encouragement over these years.

Special thanks to my committee members, Professor Ming Li, Professor Daniel Brown, Professor Brendan McConkey, and Professor Nuno Bandeira, for taking their precious time to review this thesis and provide valuable comments.

Many thanks to Professor Gilles Lajoie for his collaboration on GlycoMaster DB. I wish to sincerely thank him for offering many useful comments and suggestion on the experiments and the paper preparation.

I am thankful to Dr. Dongbo Bu, Dr. Xiaowen Liu, Dr. Shuai Cheng Li, Dr. Hao Lin, and Dr. Shiwei Sun for the knowledge and skills I gained by working with them. They provided useful discussions on both my research area and other branches of bioinformatics.

I also want to thank my friends and colleagues at University of Waterloo: Xuefeng Cui, Jiewen Wu, Wei He, Laleh Soltan Ghoraie, Bahador Khaleghi, Xianglilan Zhang, Eric Marinier, Hongnan Wang, Guangyu Feng, Richard Jang, Xiaofei Zhao, and Shiwei Li; and at Bioinformatics Solutions Inc.: Luosha Lu, Weiming Zhang, Baozhen Shan, Lei Xin, Mingjie Xie, Zefeng Zhang, Lian Yang, Weiwu Chen, and Brian Munro. They made my stay in Waterloo really joyful.

I am thankful for the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Ontario Graduate Scholarship (OGS) program, the David R. Cheriton Scholarship program, the MITACS Accelerate PhD Fellowship program, and the Faculty of Mathematics at the University of Waterloo.

Many thanks to the administrative staff for their great help during my study. In particular, I wish to thank Wendy Rush, Margaret Towell, Paula Zister, and Helen Jardine.

I am thankful to my parents-in-law for their kindly help before and after Alison came into the world. Their help provided me more time on this thesis writing.

I am grateful to my mother and father, who have always been sharing my joys and sorrows from the very beginning. They have provided their unconditional love and support at every stage of my life.

I am indebted to my dear wife and my daughter Alison for enduring my long working days and nights. It is them who make my life bright and hopeful.

Dedication

To my wife Xi, whose constant support made it possible.

Table of Contents

List of Tables	xv
List of Figures	xix
1 Introduction	1
1.1 Background	1
1.2 Contributions	2
1.3 Overview of Dissertation	4
1.4 Publication Notes	5
2 Fundamentals	7
2.1 Mass Spectrometry	7
2.2 Shotgun Proteomics	9
2.3 Tandem Mass Spectrometry	9
2.4 Peptide Fragmentation	10
2.5 Glycopeptide Fragmentation	12
2.6 Peptide Identification	13
2.7 Protein Quantification	16
3 <i>De Novo</i> Sequencing with MS/MS Spectrum Pairs	19
3.1 Methods	20
3.1.1 Peak Significance Value	21

3.1.2	Likelihood Scores of Frequent Ion Types	23
3.1.3	Score for Each Fragmentation Site	24
3.1.4	Score for Each Residue	25
3.1.5	Peptide Score	25
3.2	Experimental Results	26
3.2.1	LTQ-Orbitrap Data Set	27
3.2.2	Iontrap Data Set	28
3.3	Discussion	30
4	<i>De Novo</i> Sequencing with Many PTMs	33
4.1	Problem Formulation	34
4.2	Methods	35
4.2.1	Scoring Function	35
4.2.2	DeNovoPTM Algorithm	36
4.3	Experiments and Results	37
4.3.1	Performance Evaluation on the ISB Data Set	37
4.3.2	Performance Evaluation on the PepSplice Data Set	40
4.4	Discussion	41
5	Database Search for Modified Peptides Without Specifying PTMs	43
5.1	Methods	45
5.1.1	Protein Identification	46
5.1.2	Single-PTM Peptide Candidate Search	47
5.1.3	Modified Peptide Rescoring	47
5.1.4	Estimation of the False Discovery Rate	51
5.2	Experiments and Results	52
5.2.1	Data Sets	52
5.2.2	Coefficient Determination	52

5.2.3	Comparison between Multiple Search Engines	53
5.2.4	Comparison with MOD ⁱ	55
5.2.5	Consensus Strategy and Analysis	56
5.2.6	Summary of Identified PTMs	56
5.3	Discussion	59
6	Identification of N-linked Glycopeptides by Tandem Mass Spectrometry	61
6.1	Methods	64
6.1.1	Filtration of Glycopeptide Spectra	64
6.1.2	Glycan Assignment	66
6.1.3	Glycopeptide Identification	68
6.2	Results	69
6.2.1	RNase-B Data Set	71
6.2.2	Human-IgG Data Set	75
6.2.3	Enriched-HUP Data Set	78
6.2.4	HUP Data Set	80
6.2.5	Comparison of Identified Glycans between Enriched-HUP and HUP Data Sets	83
6.2.6	Glycopeptides with Same Mass	86
6.3	Discussion	86
7	Maximum Peptide Feature Matching in Label-Free Quantification	89
7.1	The Maximum Feature Matching Problem	92
7.2	Maximum Feature Matching Is <i>NP</i> -Hard	93
7.3	A Practical Algorithm for Maximum Feature Matching	98
7.4	Variations of the Maximum Feature Matching Problem	102
7.4.1	Weight Function	102

7.4.2	Gap Penalty	103
7.5	Experimental Results	104
7.6	Discussion	109
8	Conclusions and Future Work	113
8.1	Conclusions	113
8.2	Future Work	115
	References	136

List of Tables

2.1	Comparison of typical performance characteristics of commonly used mass analyzers in proteomics [12].	9
2.2	The 20 standard amino acid residues.	14
3.1	Frequencies of peaks matched by common fragment ion types in the training data.	23
3.2	Comparison between identifications of spectra in the LTQ-Orbitrap testing data set. The first four rows are the percentage of the peptide sequences with at most 0, 1, 2, and 3 incorrect residues in the <i>de novo</i> sequencing results. The last row is the percentage of the total correct residues.	28
3.3	Comparison of identifications on the spectra in the Iontrap testing data. The first four rows are the percentage of the peptides with at most 0, 1, 2, and 3 incorrect residues in the <i>de novo</i> sequencing results. The last row is the percentage of the total correct residues.	29
4.1	Thirty-eight PTMs used to evaluate the performance of three <i>de novo</i> sequencing software tool.	39
4.2	Comparison between the performances of three <i>de novo</i> sequencing software tools on the ISB data set. Each software tool was run three times with 4 PTMs, 38 PTMs, and 71 PTMs being specified, respectively. This table listed the numbers of PTMs that were identified correctly on both PTM types and positions, as well as the numbers of correctly identified residues.	40

4.3	Comparison between the performances of three <i>de novo</i> sequencing software tools on the PepSplice data set. Each software tool was run three times with 4 PTMs, 38 PTMs, and 71 PTMs being specified, respectively. This table listed the numbers of PTMs that were identified correctly on both PTM types, as well as the numbers of correctly identified residues.	41
5.1	The summary of 29 PTMs which are frequently reported in previous research.	50
5.2	The numbers of identified peptides with $FDR \leq 1\%$ under different settings of training and testing data sets.	53
5.3	The number of unique modified peptides containing the most common PTMs in the Human-heart data set.	58
6.1	The grouped results identified by GlycoMaster DB from the RNase-B data set. The spectra with the same precursor m/z and charge are grouped in a row if they have the same identification. The GSM score, PSM score and mass error in a row are from the HCD/ETD spectrum-pair with the highest GSM score.	73
6.2	Glycopeptide identified by GlycoMaster DB from the Human-IgG data set.	76
6.3	This table lists the analysis result of each spectral data in enriched-HUP data set. The spectra data named “gpe12” is not listed since it has no glycan reported by GlycoMaster DB. The first column lists the names of the spectral files. The second column denotes the numbers of MS/MS spectra in each file after preprocessing. The third and fourth column lists the numbers of proteins reported by PEAKS DB and the numbers of un-interpreted spectra. The subsequent two columns give the number of spectra that passed the two filters, respectively. The number of identified glycans ($-10 \lg P \geq 15$) and peptides are listed in the last columns. The last row shows the total number of each column.	79

6.4	This table lists the analysis result of each spectral data in HUP data set. Only the 23 spectral data having identified glycans by GlycoMaster DB are listed. The first column lists the names of the spectral files. The second column denotes the numbers of MS/MS spectra in each file after preprocessing. The third and fourth column lists the numbers of proteins reported by PEAKS DB and the numbers of un-interpreted spectra. The subsequent two columns give the number of spectra that passed the two filters, respectively. The number of identified glycans ($-10 \lg P \geq 15$) and peptides are listed in the last columns. The last row shows the total number of each column.	84
7.1	The number of features in different samples.	106
7.2	The comparison of average aligned time errors (in seconds) and the percentages of correctly aligned feature pairs on true peptide features.	108

List of Figures

2.1	An example of a visualized mass spectrum. The inset illustrates isotopic peaks with monoisotopic m/z 1396.14. The charge state of the ion can be determined as two from the m/z difference between two adjacent isotopic peaks.	8
2.2	A typical MS/MS experiment procedure. (1) protein digestion; (2) peptide separation; (3) survey scan generation; (4) peptide fragmentation; (5) MS/MS spectrum generation.	10
2.3	Six types of fragment ions, <i>i.e.</i> , a -, b -, c -, x -, y -, and z -ions, generated through breaking the backbone of a peptide.	11
2.4	An illustration showing the fragmentation pattern of glycopeptides.	12
2.5	An example showing an annotated mass spectrum of peptide YGFIEGHVVIPR.	15
3.1	Comparison of two approaches to incorporate peak intensities. The areas under ROC curves represent the discriminative performance of using the significance value and the relative intensity alone.	22
3.2	(a) The distributions of the peak significance for the true z' -ions and random matches. (b) The z' -ion likelihood scoring function with respect to the significance value.	24
3.3	An annotated MS/MS spectrum that has stronger b -H ₂ O ions for many adjacent fragmentation sites, indicating strong correlation between the adjacent fragment ions with the same ion type.	26

3.4	(a) Identification rates of different software as the function of the number of allowed incorrect residues for the testing data. (b) Identification rates of different software as the function of the number of allowed incorrect residues for the testing data.	30
5.1	The LDF score distributions of single-PTM peptides identified from target and decoy databases, respectively. (a) The distribution with peptide pairs, and (b) without peptide pairs. Modified peptides from the target database tend to have more peptide pairs than those from the decoy database.	48
5.2	The LDF score distributions of the peptide candidates identified with no PTM, a common PTM, and a rare PTM, from (a) the target database and (b) the decoy database.	49
5.3	The comparison of reported modified PSMs by InsPecT, Mascot, Paragon and PeaksPTM. The curves show the relation between the estimated FDR and the number of modified PSMs.	54
5.4	A large portion of modified PSMs reported by PeaksPTM with high confidence ($FDR \leq 1\%$) are also identified by at least one other engine, either with high or low confidence.	55
5.5	The comparison of PeaksPTM, MOD ⁱ and InsPecT on the reduced database with twenty proteins (ten target and ten decoy proteins). The curves show the relation between the estimated FDR and the number of modified PSMs reported.	57
5.6	The Venn diagram shows the modified PSMs reported by applying the consensus strategy on the results of four search engines.	57

- 6.1 An example of the GlycoMaster DB result page generated in an HCD/ETD spectral data analysis. The results are listed in a HTML table in descending order of glycan scores. Each row represents an identification of an HCD/ETD spectrum pair. The first column includes a hyperlink that redirects to the top-ten interpretations of the same HCD/ETD spectrum-pair. The second to the fifth column list the spectrum information. The sixth and the seventh column list the glycan information obtained from the GlycomeDB database, and the hyperlinks redirect to the GlycomeDB website. The eighth column gives scores of GSMs and each hyperlink redirects to the annotated HCD spectrum and its mass error chart. The ninth and tenth column list the peptide sequences and the PSM scores obtained from ETD spectra. The hyperlink at the PSM score column links to the annotated ETD spectrum and the mass error chart. The mass error between the theoretical and experimental mass values of an identified glycopeptide is provided in the eleventh column. The last column gives the accession numbers of corresponding proteins. 72
- 6.2 An example of a glycopeptide identified from the RNase-B data set generated by the HCD PI ETD strategy. (a) The annotated HCD spectrum of precursor ions with m/z 807.672. GlycoMaster DB identified the best matched glycan with the composition HexNAc2Hex8. SRNLTK is the only potential glycopeptide having the similar mass to the calculated mass 699.404. (b) The annotated ETD spectrum triggered by product ions in the HCD spectrum shown in (a). It provides positive support for the identification of the peptide SRNLTK and the glycosylation site. 74
- 6.3 An example of a glycopeptide identified from the Human-IgG data set generated by the HCD PI ETD strategy. (a) The annotated HCD spectrum of precursor ions with m/z 1028.7906. The best matched glycan reported by GlycoMaster DB has the composition HexNAc₄Hex₃Fuc₁. (b) The annotated ETD spectrum triggered by product ions in the HCD spectrum shown in (a). It provides positive support for the identification of peptide TKPREEQFNSTFR and the glycosylation site. . . 77

6.4	An example of glycans identified from three HCD spectra by GlycoMaster DB in the Enriched-HUP data set. Three HCD spectra have similar retention time but different precursor mass values. GlycoMaster DB identified three glycans. The calculated peptide mass is approximate 771.41. NWTITR is the only tryptic glycopeptide matching this mass value from the protein short list provided to GlycoMaster DB. The mass errors of the identifications of these three spectra are -1.29 ppm, -1.08 ppm, and -0.84 ppm, respectively.	81
6.5	Illustration of two HCD mass spectra that are interpreted as the same peptide but two slightly different glycans in the Enriched-HUP data set. (a) The oxonium ions from sialic acids are not present, and this indicates the absence of sialic acids in the glycan; (b) The peaks at m/z 292.10 and 274.09 indicate the existence of oxonium ions of sialic acid residues.	82
6.6	An example of glycans identified from three HCD spectra by GlycoMaster DB in the HUP data set. Three HCD spectra have similar retention time but different precursor mass values. GlycoMaster DB identified three glycans from them and these glycans differ from each other slightly. The calculated peptide mass is approximate 1449.74. VYKPSAGNNSLYR is one of the two peptides matching this mass value in the proteins provided to GlycoMaster DB but the other one has potassium adduct and much larger mass error at around 10 ppm. Therefore, VYKPSAGNNSLYR is selected as the glycopeptide and the precursor mass errors are 0.71 ppm, 0.08 ppm, and -0.51 ppm, respectively.	85
6.7	The Venn diagram showing the overlaps between the two sets of glycopeptide groups identified from Enriched-HUP and HUP data sets, respectively.	87
6.8	The average percentage of tryptic peptides containing the <i>N</i> -linked glycopeptide motif that have unique mass.	87
7.1	An Illustration of features plotting on a mass-time grid. Each horizontal line on the plane corresponds to one time unit in the LC-MS experiment, and each vertical line corresponds to a mass unit.	95

7.2	An illustration of the construction that highlights edge e_k and vertex v_i .	95
7.3	Illustrations of three cases to construct the grayed-out region: (a) $e_k = (v_i, v_j)$ and $i < j$; (b) $e_k = (v_j, v_i)$ and $i > j$; and (c) v_i is not adjacent to e_k .	97
7.4	A shifted matching propagates upward when $f(6i - 1) = 6i$.	97
7.5	Comparison of the feature matching software tools on data sets from different labs: iPRG vs. Coon2.F4 (a) and Coon1.F3 vs. Mann.1 (b). The x-axis denotes the retention time in the first sample and the y-axis denotes the retention time in the second sample. A blue circle, which stands for a feature pair matched according to peptide identification, is considered as the ground truth. A gray cross represents a possible feature pair matched purely by the precursor mass. The curves are produced by the compared algorithms without knowing the blue circles.	110
7.6	Comparison of the feature matching software tools on data sets from the same lab: Coon1.F3 vs. Coon2.F4 (a) and Mann.1 vs. Mann.2 (b). As msInspect and MZmine2 failed to align the data set Mann.1 and Mann.2, only results of SMFM, SMFM-g, MultiAlign, and Polynomial-4 are shown in (b).	111

Chapter 1

Introduction

1.1 Background

Proteomics refers to the comprehensive study of the entire protein content in a specific cell, tissue or organism, or body fluids, *ie.* blood and urine. Its goal is to obtain a global and integrated view of disease processes, cellular processes and networks at the protein level [18]. Qualitative and quantitative proteomic analysis can help in discovering unique biomarkers, which play extremely important roles in the diagnostic and therapeutic procedures of some diseases, such as cancer, in modern medical research [119].

Currently, tandem mass spectrometry (MS/MS) is the standard analytical technology in proteomics. It enables high-throughput analysis on proteins and their quantities in a complex protein mixture with high sensitivity, selectivity and accuracy [36]. A tandem mass spectrometer can easily and rapidly generate large volumes of spectral data for a biological sample. It makes manual interpretation become unfeasible and has also brought numerous challenges in automated data analysis. Elegant algorithmic solutions have been proposed and provide indispensable analytical support in current proteomic experiments.

Accurate identification of proteins is the preliminary requirement in proteomics. Protein sequence databases have been constructed by gathering proteins either sequenced in previous experiments or predicted from genes. MS/MS data can then be interpreted by searching these protein databases [163, 164]. In contrast, *de novo* sequencing approaches have also been developed to study un-sequenced organisms or

species, of which the proteins are directly identified from spectral data without the assistance of protein databases [34].

Inaccurate identification is the major obstacle that hinders *de novo* sequencing to be a reliable approach for protein identification [111]. Recently, instruments that integrate multiple fragmentation modules have been developed. Two or more fragmentation methods are used to generate different types of spectra from the same sample, such as collision-induced dissociation (CID), higher energy collisional dissociation (HCD), and electron-transfer dissociation (ETD). It becomes possible to improve the *de novo* sequencing by using multiple types of spectral data.

The occurrence of posttranslational modifications (PTMs) on proteins is of critical importance to protein functions [159]. Most existing software tools, using either database search or *de novo* sequencing approaches, have difficulties handling a large number of PTMs specified by users. Either the accuracy of the result or the speed of the algorithms, or both, can be seriously influenced when too many PTMs are considered [27].

N-linked glycosylation is one of the most frequently observed PTMs in mammalian organisms. The identification of glycopeptides and glycans is crucial to studies of cellular processes, particularly some disease processes [19]. Compared to peptides with other types of PTMs, intact *N*-linked glycopeptides generate spectral data with explicitly different patterns. The analysis of such data heavily depends on manual interpretation and no automated solution is currently available for a large scale analysis. This has become the major obstacle to the progress of glycoproteomics [33].

Protein quantification provides the information of protein quantity changes and assists in discovering important biomarkers of particular diseases. The label-free quantification method is one of the two commonly used approaches in protein quantification [11]. It has attracted growing interest since no additional chemistry or sample preparation steps are required. However, the convenience in experiments leads to more computational challenges, which demands efficient and accurate algorithmic solutions.

1.2 Contributions

This dissertation is mainly to propose algorithmic solutions for protein identification and PTM characterization. The following five components constitute the contribu-

tions of the dissertation:

***De novo* sequencing using CID/ETD spectrum-pairs:** The errors of *de novo* sequencing mainly come from ambiguities introduced by missing peaks. In a spectrum, incomplete information makes it difficult to discriminate amino acid combinations that share the same mass. However, in a spectrum-pair consisting of a CID spectrum and an ETD spectrum, the missing peaks in one spectrum may be present in the other. This fact helps to determine the true amino acid combination and thus improve the accuracy of *de novo* sequencing. We propose a *de novo* sequencing approach using CID/ETD spectrum-pairs, named ADEPTS. The comparison with other *de novo* sequencing software tools shows the better performance of ADEPTS.

***De Novo* sequencing with many PTMs:** Since it is difficult for a researcher to know all the PTM types in a sample, a natural practice is to consider as many PTM types as possible and let the data analysis algorithm determine which PTMs really exist. However, the accuracy of *de novo* sequencing is significantly degraded due to the larger search space introduced by considering many PTM types. We propose DeNovoPTM as a specialized application of *de novo* sequencing when many PTM types are considered. Our observation shows that most peptides in a proteomic study contain only a small number of PTMs per peptide, yet the types of PTMs can come from a large number of choices. Therefore, it is desirable to include a large number of PTM types in a *de novo* sequencing algorithm but limit the number of PTM occurrences in each peptide to increase the accuracy. A dynamic programming algorithm for solving this problem is proposed and implemented for practical use.

Searching modified peptides without specifying PTMs: Identification of modified peptides using conventional database search software tools requires users to provide a few PTM candidates in advance. However, the complete knowledge of possibly existing PTMs in a protein mixture cannot be obtained before the analysis. On the other hand, the efficiency and accuracy reduces significantly when a large number of PTMs are specified. We present an improved database search tool for modified peptide identification without pre-specifying PTM candidates. The improvements in the software include (1) a default setting whereby the software considers all PTMs included in the Unimod database as variable PTMs, and (2) several search strategies that significantly reduce the search space. Furthermore, our approach uses the co-existence of modified and base forms of the same peptide and the rareness of PTMs to provide powerful discrimination between spurious and real modified peptides. This

software outperforms several state-of-the-art software packages evaluated in this research.

Characterizing intact glycopeptides: Identification of glycopeptides and glycans is essential to better understand the functions and bioactivities of glycoproteins. The progress of this study is mainly hindered by the lack of algorithms for intact glycopeptide characterization. GlycoMaster DB is proposed to fulfil this urgent requirement on *N*-linked glycopeptides. It is able to analyze the MS/MS data obtained from a biological sample with glycoproteins being either enriched or not, and from either HCD/ETD or HCD-only fragmentation. It simultaneously identifies glycopeptide sequences and *N*-linked glycan composition from a user-specified protein database and a pre-configured *N*-linked glycan database, respectively. Furthermore, the connections between glycopeptides and glycans are also reported by GlycoMaster DB. This connection information makes it possible to determine the glycoprotein identity and study different forms of a glycoprotein (glycoforms).

Matching peptide features: Protein quantification is to study the abundance variance of interested proteins from two or more samples. For each of the proteins, its abundance ratio can be obtained by calculating the ratios of its peptides. In many cases peptides are not identified in advance and peptide features (signals that are possibly caused by peptides) are usually used to represent peptides. Thus, pairing two peptides from different samples means pairing two peptide features from different feature sets. Mass and retention time are two most important pieces of information of a peptide feature. The mass is fairly accurate, but the retention time is subject to some systematic as well as random errors. Features with the same or similar mass are matched and a matching weight according to the retention time shift is calculated. The maximum peptide feature matching problem is to compute a match between two feature sets with the maximum weight. The difficulty of this problem is finding an “alignment” that maps the retention time from one peptide feature set to the other. We formulate this problem into a combinatorial model and prove its NP-hardness. Practical algorithms are provided and compared with other existing methods.

1.3 Overview of Dissertation

The rest of the dissertation is organized as follows:

Chapter 2 introduces the fundamentals of MS/MS-based computational proteomics.

Both experimental and computational strategies are introduced for better understanding of the subsequent research topics.

Chapter 3 presents ADEPTS, a *de novo* sequencing approach that improves the accuracy by using CID/ETD spectrum-pairs.

Chapter 4 presents DeNovoPTM, a *de novo* sequencing algorithm that improves the accuracy when many PTMs are considered.

Chapter 5 proposes PeaksPTM, an improved database search approach for efficient identification of modified peptides, with the consideration of all PTMs in the Unimod database.

Chapter 6 presents GlycoMaster DB, a database search software tool for the characterization of intact *N*-linked glycopeptides from spectral data obtained by HCD/ETD or HCD-only fragmentation.

Chapter 7 studies the peptide feature matching problem encountered in label-free protein quantification. The problem is formulated into a combinatorial model and proven to be NP-hard. Two practical algorithms are provided.

Finally, the summary of this dissertation and future work are presented in Chapter 8.

1.4 Publication Notes

The studies presented in Chapter 3, 4, 5, and 7 have been published as four referred research articles [64, 63, 60, 89]. The work in Chapter 6 has been submitted for possible publication.

Chapter 2

Fundamentals

2.1 Mass Spectrometry

Mass spectrometry (MS) is an analytical technique that measures the mass-to-charge ratios (m/z) of charged particles. It has been used for both qualitative and quantitative analysis, which includes identifying the composition and structures of unknown compounds and measuring the quantities of interested molecules. Currently, MS is widely used in analytical laboratories where physical, chemical, or biological properties of a great variety of compounds are studied [168].

A mass spectrometer typically consists of three components: an ionizer, a mass analyzer, and a detector. Molecules are first converted to ions in the ionizer, then the ions are separated according to their different m/z in the mass analyzer. The separated ions are then detected by the detector to form a mass spectrum, which consists of a list of peaks. Each peak is represented by its m/z and intensity. Figure 2.1 illustrates an example of a mass spectrum. Ions with the same m/z form a peak and the intensity of a peak indicates the number of such ions detected by the detector. The isotopes of elements in a molecule also produce isotopic peaks, from which the charge state of the ion can be determined. The intensity of a peak is related to the abundance of the corresponding ion. However, the abundance ratio between two molecules cannot be simply regarded as the intensity ratio of their corresponding peaks [96]. This is due to the differences in ionization efficiency and detectability of different molecules, as well as the imperfect reproducibility of MS experiments.

The ionizer and the mass analyzer of a mass spectrometer are implemented with

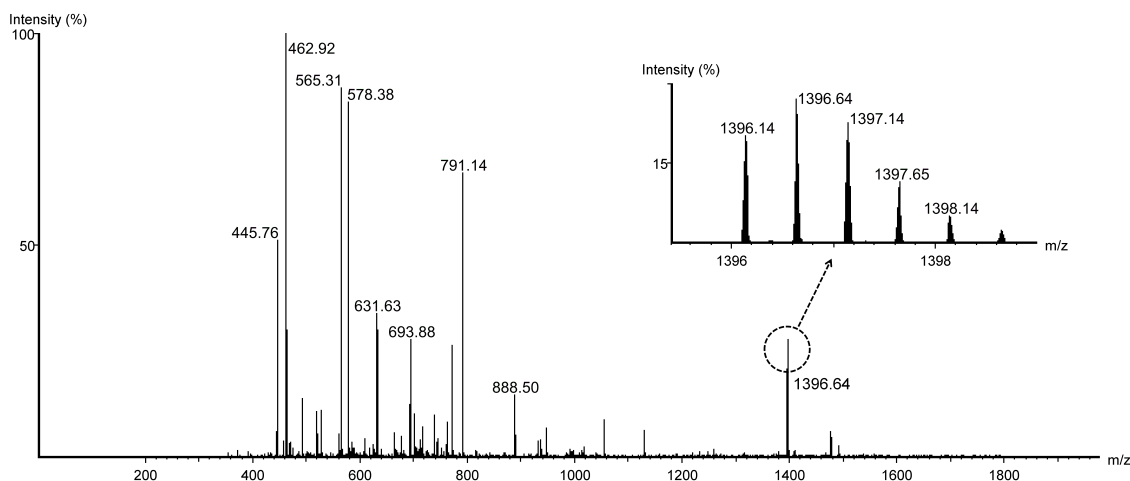


Figure 2.1: An example of a visualized mass spectrum. The inset illustrates isotopic peaks with monoisotopic m/z 1396.14. The charge state of the ion can be determined as two from the m/z difference between two adjacent isotopic peaks.

multiple techniques, causing different properties of the resultant spectral data. Two types of ionizers are commonly used in proteomics: matrix-assisted laser desorption/ionization (MALDI) [66] and electrospray ionization (ESI) [102]. MALDI mostly produces singly charged ions, and ESI can produce multiply charged ions. The advantage of ESI is that a large molecule can still be detected since its multiply charged ions can fall into the m/z range of a mass spectrometer. However, the existence of multiply charged ions increases the complexity of the spectrum and more computational efforts are required for the spectrum interpretation.

Five types of mass analyzers are commonly used in proteomics: quadrupole, ion trap (quadrupole ion trap, QIT; linear ion trap, LIT or LTQ), time-of-flight (TOF), Fourier transform ion cyclotron resonance (FTICR), and Orbitrap. Each type has different capabilities in terms of sensitivity, accuracy, resolution, m/z range, and other characteristics [68]. Mass resolution and accuracy are two important parameters to measure the performance of a mass analyzer. The mass resolution measures the ability to distinguish two peaks of slightly different m/z . The mass accuracy is the ratio between the m/z measurement error and the true m/z , and usually measured in ppm (parts per million). The performance of each mass analyzer is listed in Table 2.1.

Table 2.1: Comparison of typical performance characteristics of commonly used mass analyzers in proteomics [12].

Mass analyzer	Resolution	Accuracy (ppm)	m/z range	Scan rate
Quadrupole	1,000	100-1,000	50-2,000; 200-4,000	Moderate
QIT	1,000	100-1,000	10-4,000	Moderate
LTQ	2,000	100-500	50-2,000; 200-4,000	Fast
TOF	10,000-20,000	10-100	No upper limit	Fast
FTICR	100,000-750,000	<2	50-2000; 200-4,000	Slow
Orbitrap	30,000-100,000	2-5	50-2,000; 200-4,000	Moderate

2.2 Shotgun Proteomics

An explicit goal of proteomics is to characterize all the proteins expressed in a cell or tissue [4]. The improvements in mass spectrometry instruments, protein and peptide separation techniques, and the availability of protein sequence databases for many species has facilitated the analysis of complex protein mixtures using shotgun proteomics. The major steps of shotgun proteomics include the protein digestion by single or multiple enzymes, the peptide separation by liquid chromatography (LC), and the peptide analysis using tandem mass spectrometry technology. Peptides are identified from the spectral data, and proteins can then be determined by matching these identified peptides to known protein sequences or assembling them into novel proteins [2, 105]. Shotgun proteomics is currently the dominant analytical approach in proteomics research [168].

2.3 Tandem Mass Spectrometry

Tandem mass spectrometry (MS/MS) technology involves multiple mass spectrometer stages and aims to precisely identify and characterize peptide sequences. A typical MS/MS-based proteomic experiment contains the following steps, as illustrated in Figure 2.2: (1) protein digestion to produce shorter peptides; (2) peptide separation by LC or other separation approaches; (3) analysis and detection of peptide ions; (4) peptide selection for further fragmentation; (5) analysis and detection of peptide

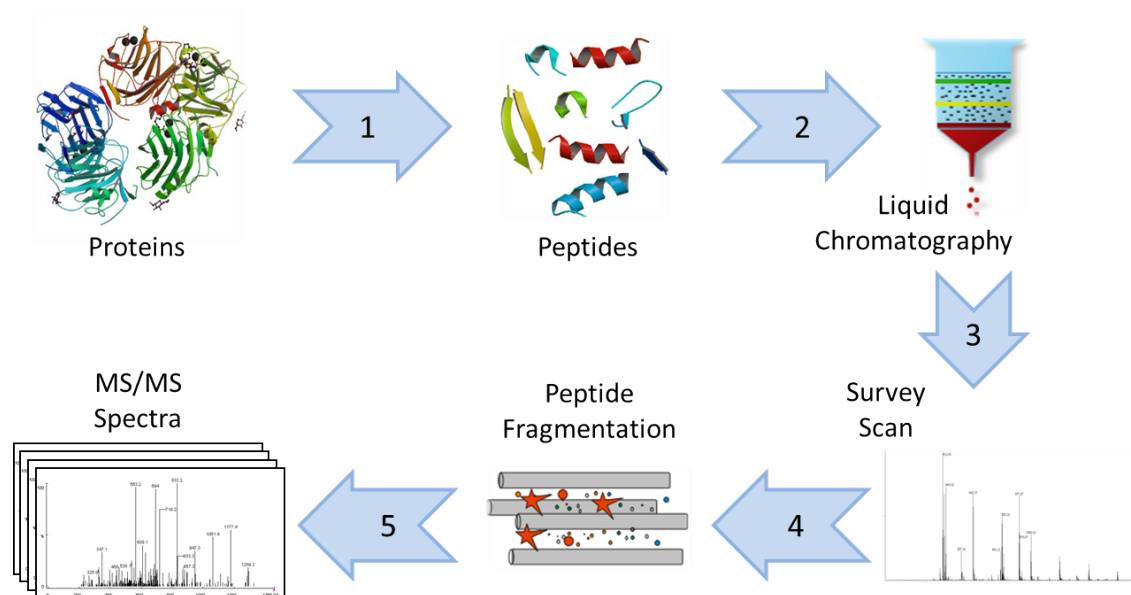


Figure 2.2: A typical MS/MS experiment procedure. (1) protein digestion; (2) peptide separation; (3) survey scan generation; (4) peptide fragmentation; (5) MS/MS spectrum generation.

fragment ions [168]. Two types of mass spectra are generated in such an MS/MS experiment: *survey scans* and MS/MS spectra. Each survey scan has a *retention time*, indicating the time in the LC experiment when the survey scan is taken. Each peak in a survey scan denotes a peptide ion. Its m/z value reflects the mass of the whole peptide but can not provide more information of the peptide sequence. The subsequent fragmentation breaks a selected peptide to generate a series of fragment ions, of which the m/z values are recorded in an MS/MS spectrum. Analysis of MS/MS spectra, sometimes with the assistance of corresponding survey scans, can help to identify peptide sequences and then determine the proteins in a biological sample.

2.4 Peptide Fragmentation

In an MS/MS experiment, peptides are further fragmented to identify the sequences. Peaks in an MS/MS spectrum represent a set of fragment ions generated from the dissociation of a selected peptide. Theoretically, the peptide backbone can be broken

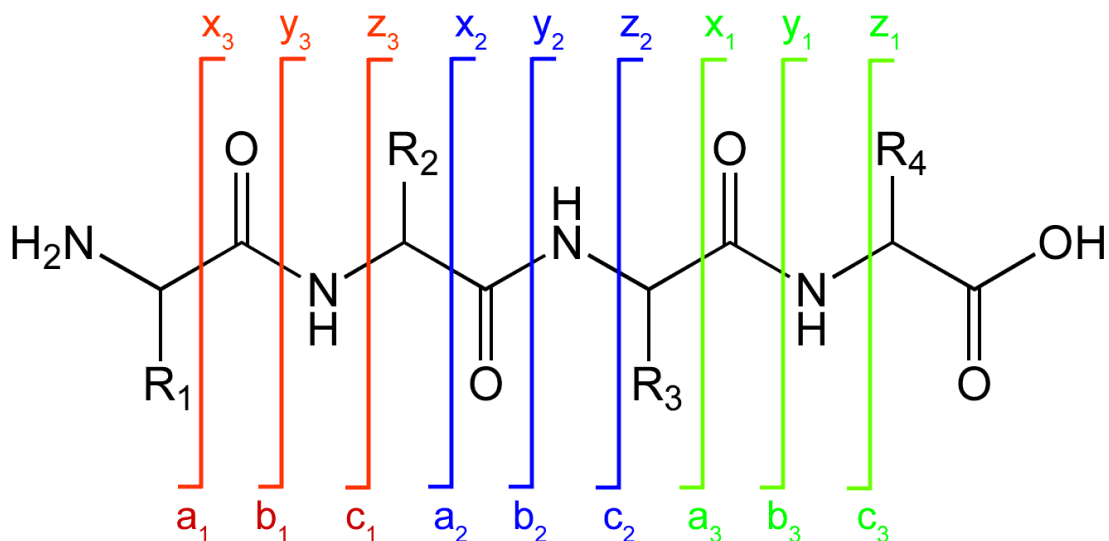


Figure 2.3: Six types of fragment ions, *i.e.*, a -, b -, c -, x -, y -, and z -ions, generated through breaking the backbone of a peptide.

at any of three sites per residue to generate six types of fragment ions, as shown in Figure 2.3.

Different fragmentation approaches emphasize the generation of different ion types. Three fragmentation methods are commonly used in current MS/MS-based proteomics: In *collision-induced dissociation* (*CID*), also known as *collisionally activated dissociation* (*CAD*) [25], the peptide ions are usually accelerated by some electrical potential to high kinetic energy and then collided with neutral molecules (often helium, nitrogen or argon). In the collision some of the kinetic energy is converted into internal energy which results in bond breakage. Two types of fragment ions, b - and y -ions, are frequently observed in MS/MS spectra obtained by CID fragmentation. In *higher energy collisional dissociation* (*HCD*) [114], peptide ions are injected into a collision cell and fragment ions are then analyzed by an Orbitrap analyzer. The mechanism of HCD is similar to CID but more accurate m/z for the fragment ions can be measured. HCD also generates b - and y -ions dominantly. When the higher energy is deployed, b -ions can be further fragmented into a -ions or smaller species. Lastly, *electron-transfer dissociation* (*ETD*), or *electron-capture dissociation* (*ECD*) [143], transfers electrons to a multiply protonated peptide/protein, or generates radical cations for a multiply protonated peptide/protein, and leads to the cleavage of the $N-C_\alpha$ backbone bonds

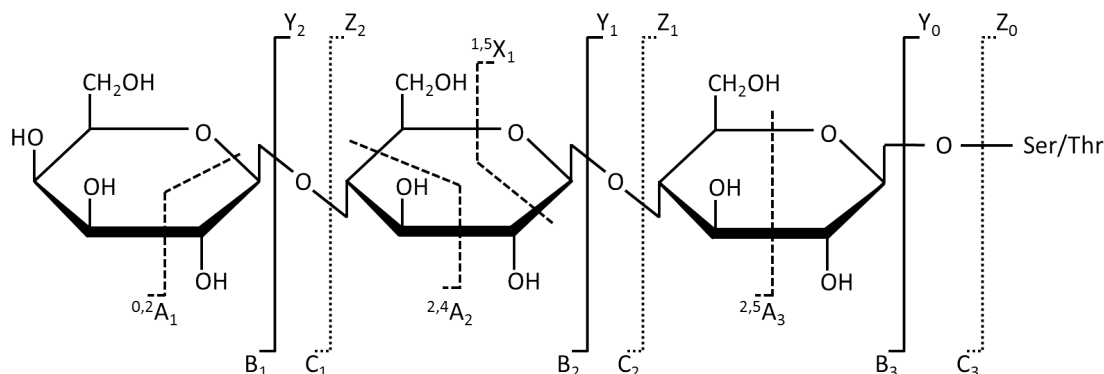


Figure 2.4: An illustration showing the fragmentation pattern of glycopeptides.

to generate c - and z -ions. Ions derived from these ions, such as $(c-1)$ - and z' -ions, are also observed frequently in ETD (or ECD) spectral data [106, 171].

2.5 Glycopeptide Fragmentation

Glycosylation can decrease the fragmentation efficiency of peptide backbones when collision based dissociation methods are used. For instance, fragmentation patterns of glycopeptides are different from non-glycosylated peptides when using CID or HCD. The collision energy is mainly absorbed by glycans. This leads to the glycosidic bond breakages, while the peptide bonds are seldom broken. Thus, the fragment ions generated by such approaches are dominantly B -, Y - and oxonium ions, as well as some cross-ring fragment ions (A - and X -ions). The notations of fragment ions generated from glycosidic bonds are different from the ones from peptide bonds (using lower cases, *e.g.*, b - and y -ions). The introduction of these notations can be found in Domon and Costello nomenclature [37]. The breakages of glycosidic bonds in CID or HCD fragmentation are illustrated in Figure 2.4. In contrast, ETD and ECD dominantly produce fragment ions by breaking a peptide backbone but retain the attaching glycan. Thus, glycans can be readily treated as normal PTMs with large mass deviations in such case. The dominant types of fragment ions generated by ETD (or ECD) for glycopeptides are c -, z -ions and their derived ions.

2.6 Peptide Identification

Peptides are identified through interpreting MS/MS spectra data based on the prior knowledge of common amino acids and PTMs. The 20 common amino acid residues are listed in Table 2.2. The interpretation of an MS/MS spectrum seeks the best matching peptide for the given spectrum. For example, in Figure 2.5, the peptide YGFIEGHVVIPR is the best interpretation of the spectrum and its theoretical y -ions generated from this peptide match all the significant peaks in the spectrum with small mass errors.

In general, a *peptide-spectrum match* (*PSM*) score is calculated to measure the similarity between a peptide candidate and a spectrum. A *fragmentation site* refers to all types of ions generated from a fragmentation between two adjacent amino acid residues. The peaks matched by a fragmentation site in the spectrum are used to calculate a score for this fragmentation site, and the PSM score of a peptide candidate is calculated from the combination of the scores at all the fragmentation sites. The peptide candidate with the highest PSM score is finally selected as the identification of the spectrum. The approach to calculate a PSM score is called a *scoring function*, or a *scoring scheme*, which is the core part of the whole procedure of peptide identification from a spectrum.

Database search and *de novo* sequencing are the two mainly used computational approaches for spectral data interpretation. The major difference of these two approaches is the requirement of protein databases.

A database search approach requires the assistance of protein databases. The protein sequences in a protein database are digested *in silico* to generate peptides, then an MS/MS spectrum is compared with each possible peptide to calculate a PSM score. The identification of the MS/MS spectrum is reported as the peptide from the database with the top PSM score. The popular database search software packages include Mascot [118], PEAKS [166], Sequest [41], MS-GFDB [78], X!Tandem [28], and OMSSA [52].

In contrast, *de novo* sequencing constructs the peptide sequence directly from an MS/MS spectrum, thus it is often used for novel protein identification. Rather than searching peptides from a protein database, it searches all amino acid combinations to find the optimal peptide sequence. Such searching is usually carried out by an efficient dynamic programming algorithm to avoid the exponential running time. PEAKS

CHAPTER 2. FUNDAMENTALS

Table 2.2: The 20 standard amino acid residues.

Name	3-letter Symbol	1-letter Symbol	Monoisotopic Mass	Residue Composition	Residue Structure
Alanine	Ala	A	71.037114	C ₃ H ₅ NO	$\begin{array}{c} \text{CH}_3 \\ \\ \text{-NH-CH-CO-} \end{array}$
Arginine	Arg	R	156.101111	C ₆ H ₁₂ N ₄ O	$\begin{array}{c} \text{-CH}_2\text{-(CH}_2\text{)}_2\text{-NH-C-NH}_2 \\ \qquad \qquad \qquad \\ \text{-NH-CH}_2\text{-CO-} \qquad \text{NH} \end{array}$
Asparagine	Asn	N	114.042927	C ₄ H ₆ N ₂ O ₂	$\begin{array}{c} \text{CH}_3\text{-CONH}_2 \\ \\ \text{-NH-CH-CO-} \end{array}$
Aspartic Acid	Asp	D	115.026943	C ₄ H ₅ NO ₃	$\begin{array}{c} \text{CH}_3\text{-COOH} \\ \\ \text{-NH-CH-CO-} \end{array}$
Cysteine	Cys	C	103.009185	C ₃ H ₅ NOS	$\begin{array}{c} \text{CH}_2\text{-SH} \\ \\ \text{-NH-CH-CO-} \end{array}$
Glutamic Acid	Glu	E	129.042593	C ₅ H ₇ NO ₃	$\begin{array}{c} \text{CH}_2\text{-CH}_2\text{-COOH} \\ \\ \text{-NH-CH-CO-} \end{array}$
Glutamine	Gln	Q	128.058578	C ₅ H ₈ N ₂ O ₂	$\begin{array}{c} \text{CH}_2\text{-CH}_2\text{-CONH}_2 \\ \\ \text{-NH-CH-CO-} \end{array}$
Glycine	Gly	G	57.021464	C ₂ H ₃ NO	$\text{-NH-CH}_2\text{-CO-}$
Histidine	His	H	137.058912	C ₆ H ₇ N ₃ O	$\begin{array}{c} \text{CH}_2\text{---} \begin{array}{c} \diagup \text{N} \\ \diagdown \text{H} \end{array} \\ \\ \text{-NH-CH}_2\text{-CO-} \end{array}$
Isoleucine	Ile	I	113.084064	C ₆ H ₁₁ NO	$\begin{array}{c} \text{CH(CH}_3\text{)-CH}_2\text{-CH}_3 \\ \\ \text{-NH-CH-CO-} \end{array}$
Leucine	Leu	L	113.084064	C ₆ H ₁₁ NO	$\begin{array}{c} \text{CH}_2\text{-CH(CH}_3\text{)}_2 \\ \\ \text{-NH-CH-CO-} \end{array}$
Lysine	Lys	K	128.094963	C ₆ H ₁₂ N ₂ O	$\begin{array}{c} \text{CH}_2\text{-(CH}_2\text{)}_3\text{-NH}_2 \\ \\ \text{-NH-CH-CO-} \end{array}$
Methionine	Met	M	131.040485	C ₅ H ₉ NOS	$\begin{array}{c} \text{CH}_2\text{-CH}_2\text{-S-CH}_3 \\ \\ \text{-NH-CH-CO-} \end{array}$
Phenylalanine	Phe	F	147.068414	C ₉ H ₉ NO	$\begin{array}{c} \text{CH}_2\text{-} \begin{array}{c} \diagup \\ \diagdown \end{array} \\ \\ \text{-NH-CH-CO-} \end{array}$
Proline	Pro	P	97.052764	C ₅ H ₇ NO	$\begin{array}{c} \diagup \text{N} \\ \diagdown \text{H} \\ \text{-N-CH-CO-} \end{array}$
Serine	Ser	S	87.032028	C ₃ H ₅ NO ₂	$\begin{array}{c} \text{CH}_2\text{-OH} \\ \\ \text{-NH-CH-CO-} \end{array}$
Threonine	Thr	T	101.047679	C ₄ H ₇ NO ₂	$\begin{array}{c} \text{CH(OH)-CH}_3 \\ \\ \text{-NH-CH-CO-} \end{array}$
Tryptophan	Trp	W	186.079313	C ₁₁ H ₁₀ N ₂ O	$\begin{array}{c} \text{-CH}_2\text{---} \begin{array}{c} \diagup \text{N} \\ \diagdown \text{H} \end{array} \\ \\ \text{-NH-CH}_2\text{-CO-} \end{array}$
Tyrosine	Tyr	Y	163.06332	C ₉ H ₉ NO ₂	$\begin{array}{c} \text{-CH}_2\text{-} \begin{array}{c} \diagup \\ \diagdown \end{array} \text{-OH} \\ \\ \text{-NH-CH}_2\text{-CO-} \end{array}$
Valine	Val	V	99.068414	C ₅ H ₉ NO	$\begin{array}{c} \text{CH(CH}_3\text{)}_2 \\ \\ \text{-NH-CH-CO-} \end{array}$

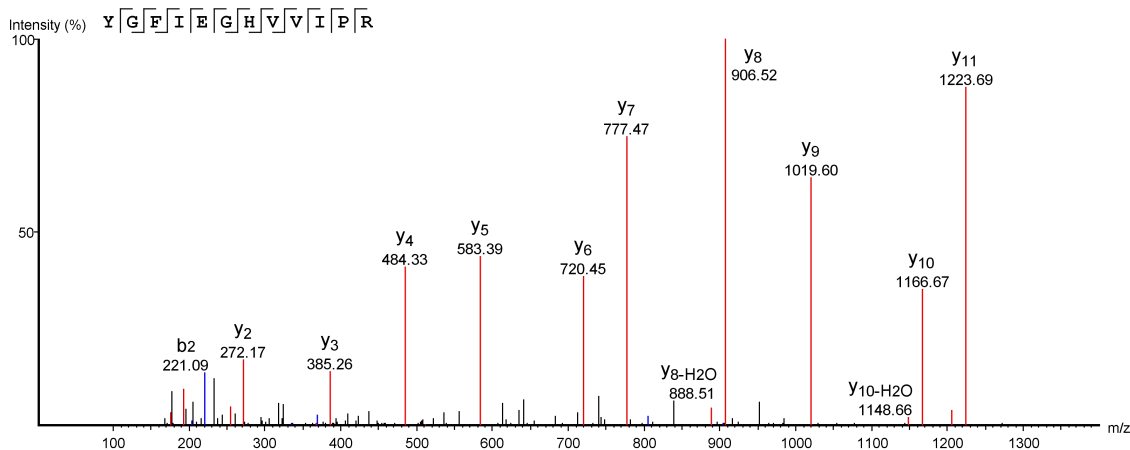


Figure 2.5: An example showing an annotated mass spectrum of peptide YGFIEGHVVIPR.

and PepNovo [45] are two state-of-the-art *de novo* software tools. Accurate *de novo* sequencing, accompanied by different enzymatic digestions, has been proven to be able to calculate the whole sequence for a purified protein sample [90]. However, the accuracy obtained from *de novo* sequencing approaches is often lower than the one obtained from database search approaches, thus there still exists much space for the improvement of *de novo* sequencing.

Spectral library search is another type of computational approach for peptide identification [29, 48, 83]. It requires a library of spectra that have been identified previously. A spectrum is then searched in the library to find the most similar counterpart, of which the corresponding peptide is reported as the search result. The discussion of this method is out of the scope of this dissertation.

A major challenge in protein identification using these computational approaches is introduced by the ubiquitous incorporation of hundreds of PTMs [39]. Most eukaryotic proteins are posttranslationally modified [159] and biochemists believe that PTMs of a protein can help determine its activity state, localization, turnover, and interactions with other proteins [100]. Therefore, precisely identifying modified proteins and their PTM types, as well as locating the modification sites, are essential to thoroughly understand their biological functions [8, 39, 100, 158]. So far, two common PTM databases, DeltaMass [1] and Unimod [31], have recorded more than 300 and 600 types of PTMs, respectively.

False positives unavoidably exist in the identification results because of the imperfect data and scoring functions. Researchers often use a *false discovery rate (FDR)* to measure the error rate in the result. The *target-decoy approach* has been widely used to validate the result by estimating the FDR [40]. In such a method, a random database (the *decoy database*) is generated with similar statistical properties as the target database. A database search approach is performed on both the target and decoy databases. The FDR at a given score threshold is then estimated by the number of matches in the decoy database with scores above the threshold. Generally, the decoy database is constructed by reversing the protein sequences in a target database, but there is still no consensus in this community on the optimal way of using decoy database. The target-decoy approach is doubted because of the pitfalls and dangers in its applications [23, 56]; however, it is still prevalently used by researchers cautiously and a modified target-decoy approach has also been proposed for two-pass database search strategies [15].

2.7 Protein Quantification

Quantitative analysis of the proteins in a cell or tissue is another important application in life science. After the identification of proteins in a sample, the expression level of each protein can help to reveal more information about the protein's participation in a particular function or malfunction of the cell [65]. Protein quantification (also known as *quantitation*) can provide a comprehensive description of the expression level changes of the proteins under the influence of various perturbations, including stress, infection, or disease. It can help identify biomarkers of particular diseases and aid in an early diagnosis and intervention. Drug administration and therapeutic effects could also be determined through protein quantification [116].

In an LC-MS experiment for peptide quantification, the peptides in a complex sample are separated by LC according to their hydrophobicity and eluted at different *retention time*. The m/z values of the co-eluting peptides are then measured by MS and a mass spectrum (survey scan) at each scanned retention time is produced. Many *peptide features* can be detected from a spectral dataset [165]. Each correctly detected feature corresponds to a peptide in the sample and mainly consists of three pieces of information: the mass, the retention time, and the signal intensity. For the same peptide, the signal intensity is approximately proportional to the abundance of such

peptide in the sample. Thus, if two features of the same peptide from two samples are confidently matched, the quantity change of the peptide can be estimated from the intensity ratio of the two matched features. The protein ratios can then be calculated from corresponding peptide quantity changes.

Two major experimental approaches exist for peptide quantification: isotopic labeling and label-free [149]. In the isotopic labeling approach, two samples are labeled with different isotopic reagents before being mixed together and then analyzed in a single LC-MS experiment. Most commercially available labeling reagents do not influence the retention time of a peptide. The same peptide with different labels from the two samples appear at almost the same retention time, making the match finding computationally simple. Labeling quantification is not covered in this dissertation and its detailed introduction is available in the literature [57, 115, 157]. The label-free method does not label the samples and measures the two samples in separate LC-MS runs. As the isotopic labeling step is not needed, the complexity of the experiment is greatly reduced [113, 156]. Label-free quantification is becoming the most promising method for the large-scale comparison of hundreds of samples that are required for biomarker discovery [169].

Label-free quantification provides more computational challenges to bioinformaticians. The major problems encountered in a label-free quantification analysis include peptide feature detection and matching, peptide ratio calculation, and protein ratio calculation. In this dissertation, the peptide feature matching problem is addressed.

Chapter 3

De Novo Sequencing with MS/MS Spectrum Pairs

De novo sequencing is important for novel protein identification in MS/MS-based proteome analysis. Nevertheless, current scientists admit that the peptides identified by *de novo* sequencing are not as confident as those from database search approaches. Higher quality MS/MS spectral data are generally required to obtain satisfactory results in *de novo* sequencing. If some of the expected fragment ions are not produced by the fragmentation, the corresponding peaks will be absent from the spectrum, leading to ambiguity for the determination of some local segments of the peptide. The ambiguity can often be eliminated in database searching, while it remains in *de novo* sequencing and leads to a partially correct peptide. Thus, more information is demanded to improve the identification accuracy, and it is essential to make *de novo* sequencing more practical in proteomics research.

Development of various fragmentation approaches provides the possibility of using multiple fragmentation methods to reduce the ambiguity in *de novo* sequencing. When a fragment ion is absent from one fragmentation method, a different type of fragment ion corresponding to the same segment of residues may be produced in another fragmentation, helping to retrieve the local peptide composition without ambiguity. In addition, peaks from the multiple spectra can confirm each other, and this can greatly increase the confidence in distinguishing a signal peak from noise. Indeed, three *de novo* sequencing approaches have benefitted from the use of two types of spectra acquired from different fragmentation methods [129, 35, 16].

Savitski *et al.* [129] introduced the first *de novo* sequencing approach using two complementary fragmentation techniques, CAD and ECD. In their method, a series of simple criteria were used to determine the correct fragmentation sites from both types of spectra. Datta *et al.* [35] computed the score of a fragmentation site by using a TAN-structured Bayesian network which involved different fragment ions from two types of spectra. This Bayesian network differs from the one used in PepNovo [45] algorithm, in which the structure is hand-selected by a human expert. The advantage of using a Bayesian network is that the correlation between different fragment ion types is considered, but the intensity information has to be discarded since the events in a Bayesian network are required to be discrete. Bertsch *et al.* [16] utilized peak intensity in their scoring scheme: theoretical spectra were generated from peptide candidates and then compared with the experimental spectrum. An accurate predictor of the theoretical spectrum is required and the performance of such a scoring scheme heavily depends on the accuracy of the spectrum predictor. However, the implementation of such a spectrum predictor is extremely difficult because of the complicated dissociation mechanism [96].

New *de novo* sequencing algorithms have been presented based on two types of spectra in all these three approaches. Nevertheless, most existing *de novo* sequencing algorithms [34, 45, 97] can also be easily applied on multiple types of spectra by simply modifying their scoring functions. In this chapter, we propose a novel scoring module, ADEPTS, to improve *de novo* sequencing performance by using CID/ETD spectrum pairs. ADEPTS uses two models to incorporate peak intensity and ion type information into the score, respectively. We show that ADEPTS increases the ability to distinguish the true fragmentation sites from the false ones and finally improves the result accuracy.

3.1 Methods

ADEPTS accepts a CID/ETD spectrum pair and a peptide candidate to calculate their matching score. Different from traditional *de novo* sequencing approaches that work only on spectra of a single type, ADEPTS uses both spectra when calculating the score for a peptide candidate. ADEPTS is used in the following framework to identify a spectrum pair:

1. Use the *de novo* sequencing module of PEAKS [97] to generate 1,000 peptide

-
- candidates for the CID and ETD spectrum, respectively;
2. Use ADEPTS to evaluate the match between each of the 2,000 peptide candidates and the spectrum pair;
 3. Report the peptide candidate with the top matching score.

The PEAKS *de novo* sequencing module in the first step can be replaced by any other existing *de novo* sequencing software that can output multiple peptide candidates from one spectrum, such as Lutefisk [148] or the algorithm proposed by Lu and Chen [94]. The second step, in which the match between a peptide candidate and a spectrum pair is evaluated, is the essential part of ADEPTS and briefly described as follows:

1. Each peak in the two spectra is assigned a non-negative *significance* value;
2. For each peptide candidate:
 - (a) Calculate theoretical m/z values of each fragmentation site and match them to peaks in the corresponding spectrum. The significance values of matched peaks form a significance vector;
 - (b) Calculate a likelihood score vector from the significance vector, then use a support vector regression (SVR) model to convert the score vector to a score for the fragmentation site;
 - (c) Similar to 2(b), calculate an SVR score for each residue of the peptide candidate;
 - (d) Add the scores for all fragmentation sites and all residues of a peptide candidate to calculate the peptide score.

Each step is described in details in subsequent sections.

3.1.1 Peak Significance Value

To calculate the significance value of a peak in an MS/MS spectrum, four features of the peak, the *rank*, the *relative intensity*, the *local rank*, and the *local relative intensity*, are considered. This approach was proposed by Liu *et al* [91]. The *rank*

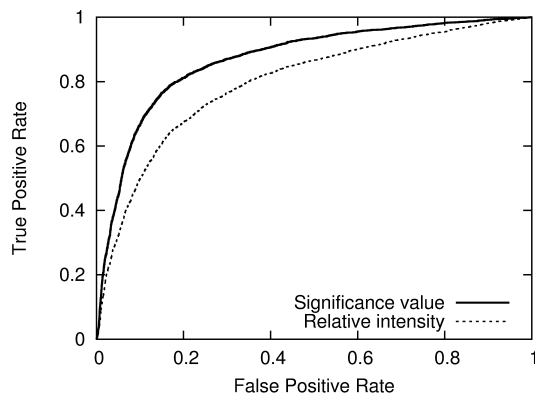


Figure 3.1: Comparison of two approaches to incorporate peak intensities. The areas under ROC curves represent the discriminative performance of using the significance value and the relative intensity alone.

of a peak is the number of peaks in the spectrum with higher or the same intensity. The *relative intensity* is the ratio between the average intensity of the top few peaks in the spectrum and the intensity of the examined peak. The definition of the *local rank* and the *local relative intensity* are the same as the rank and the relative intensity respectively, except that only the peaks within ± 56 Da from the examined peak are considered rather than all the peaks in the spectrum. The *peak significance value* S is defined as a linear combination of the logarithms of these four values:

$$S = c_1 \cdot \lg(R) + c_2 \cdot \lg(I) + c_3 \cdot \lg(R_l) + c_4 \cdot \lg(I_l), \quad (3.1)$$

where R , I , R_l , and I_l denote the rank, the relative intensity, the local rank and the local relative intensity, respectively. The coefficients ($c_i, 1 \leq i \leq 4$) are trained from an annotated training data set. According to this definition, a smaller significance value indicates a stronger peak.

We preprocess the peaks in a spectrum, including centroiding, de-isotope and de-convolute, using the data refine module of PEAKS and calculate the significance value of each peak using Eq. 3.1. Receiver operating characteristic (ROC) analysis is adopted to obtain the four coefficients c_1 , c_2 , c_3 , and c_4 . To generate the ROC curves, the peaks matched by y - and b -ions (in CID spectra) or by c - and z' -ions (in ETD spectra) are positives, and the peaks matched by randomly generated m/z values are negatives. Changing the significance threshold with a minor step can generate a series of true positive rate and false positive rate pairs, which are plotted as a ROC curve.

Table 3.1: Frequencies of peaks matched by common fragment ion types in the training data.

CID	Ion Type	y	b	$y\text{-H}_2\text{O}$	$b\text{-H}_2\text{O}$	$b\text{-NH}_3$
	Frequency (%)	52.4	47.3	31.9	31.7	27.6
	Ion	a	$y\text{-NH}_3$	$a\text{-NH}_3$	$z\text{-H}_2\text{O}$	$a\text{-H}_2\text{O}$
	Frequency (%)	25.3	23.7	23.1	22.7	22.5
ETD	Ion Type	c	y	z'	$z'+1$	$c\text{-NH}_3$
	Frequency (%)	56.6	50.4	42.6	32.4	31.6
	Ion	$y\text{-H}_2\text{O}$	z	$c\text{-H}_2\text{O}$	x	$z'\text{-NH}_3$
	Frequency (%)	20.9	20.8	19.5	18.7	18.5

A larger area under the ROC curve (AUC) denotes a better discriminative power. The four coefficients are trained using our training data to maximize the AUC.

Relative intensity is commonly used in scoring functions proposed in previous research. Figure 3.1 illustrates a comparison on the discriminative power between using the significance value and the relative intensity. It shows that the significance value, which includes the other three terms of Eq. 3.1, performs much better than relative peak intensity.

3.1.2 Likelihood Scores of Frequent Ion Types

Frequencies of peaks matched by several common ion types in our training data are listed in Table 3.1. The most frequently observed fragment ion types induced by CID and ETD are considered in ADEPTS: y , b , $y\text{-H}_2\text{O}$, $b\text{-H}_2\text{O}$, and $b\text{-NH}_3$ ions for CID spectra; c , y , z' , $z'+1$ and $c\text{-NH}_3$ ions for ETD spectra.

A likelihood score function $f_t(\cdot)$ is defined for each selected ion type t , and $f_t(x)$ represents the score of a match between a type- t ion and a peak with a significance value x .

The significance values of the matched peaks by type- t ions in the training data are divided into four intervals, each of which contains the same number of the matched peaks. The likelihood score at the centroid of each interval I_i , denoted by x_o^i , is then

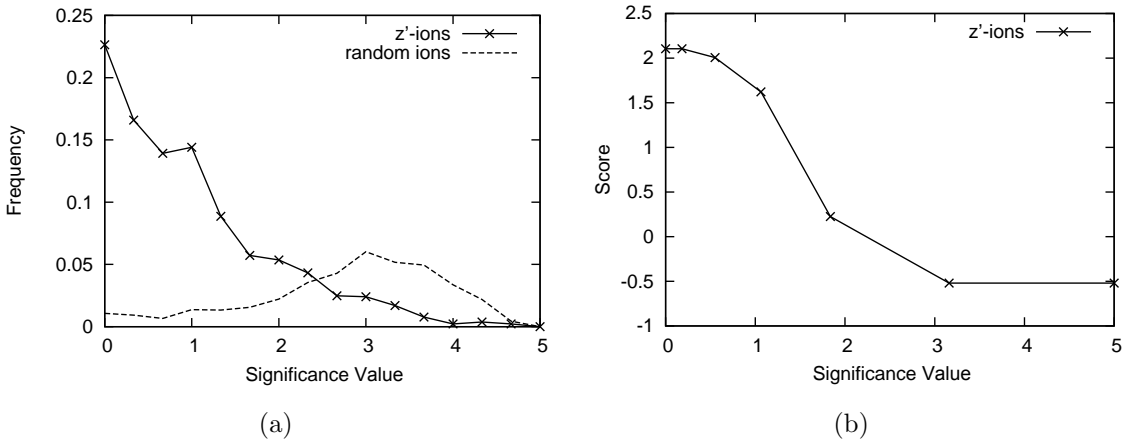


Figure 3.2: (a) The distributions of the peak significance for the true z' -ions and random matches. (b) The z' -ion likelihood scoring function with respect to the significance value.

calculated as

$$f_t(x_o^i) = \log \left(\frac{\Pr(\text{significance value falls in } I_i \mid \text{true site})}{\Pr(\text{significance value falls in } I_i \mid \text{random site})} \right). \quad (3.2)$$

The likelihood scores of significance values other than the centroid are computed by linear interpolation. If a theoretical ion does not match any peak in the spectrum, the likelihood score is calculated by

$$f_t(\text{null}) = \log \left(\frac{\Pr(\text{no peak matching} \mid \text{true site})}{\Pr(\text{no peak matching} \mid \text{random site})} \right). \quad (3.3)$$

For example, the distributions of the significance of z' -ion and random matches, as well as the likelihood scoring function of the z' -ions, are illustrated in Figure 3.2.

3.1.3 Score for Each Fragmentation Site

A peptide candidate with n amino acid residues contains $n - 1$ fragmentation sites. For each site i and ion type t , let the significance value of the matched peak be $x_{i,t}$. Then the likelihood score of this match is $s_{i,t} = f_t(x_{i,t})$, where f_t is defined in Section 3.1.2. The score of the fragmentation site i , denoted as s_i , is defined as the linear combination of likelihood score of all ion types at this site,

$$s_i = \sum_{t=1}^k c_t \cdot s_{i,t}, \quad (3.4)$$

where k is the number of selected ion types, and c_1, \dots, c_k are constant coefficients that are trained by an SVR model with the linear kernel.

The standard LIBSVM library [24] is used to train the SVR model. The training of the coefficients maximizes the distinction between the fragmentation sites of true peptide sequences and randomly generated peptide sequences using the training data.

3.1.4 Score for Each Residue

Each residue in the midst of a peptide sequence introduces two fragmentation sites at both sides of the residue. In an MS/MS spectrum, there are often strong correlations between the ions from two adjacent fragmentation sites. Figure 3.3 shows an example of such a correlation. A residue score is thus used to incorporate this correlation into the scoring function. For a specific ion type t , let x_i and x_{i+1} be the significance values of the two peaks at the two sites determined by a residue r , then the *residue significance value* x_r for this residue r and ion type t is defined as

$$x_{r,t} = \sqrt{(x_i^2 + x_{i+1}^2)/2}. \quad (3.5)$$

Clearly, if $x_i + x_{i+1}$ is fixed, the smallest $x_{r,t}$ (smaller means more significant) is achieved when $x_i = x_{i+1}$. Eq. 3.5 reflects that a residue in the midst of a peptide candidate tends to be a correct one if it determines two same-type fragment ions that match two peaks with similar significance values. This property results that a residue having two adjacent ions with similar significance values obtains a higher likelihood score. This score is used to represent the aforementioned correlation between ions.

Using the same procedure introduced to score a fragmentation site, the residue significance values of different ion types form a likelihood score vector and then this score vector is converted to a residue score s^r using an SVR model.

3.1.5 Peptide Score

Given a peptide candidate P with n amino acid residues, let the scores calculated for the $n - 1$ fragmentation sites be s_1, \dots, s_{n-1} , and the scores calculated for the $n - 2$ non-boundary residues of the peptide be s_1^r, \dots, s_{n-2}^r . The peptide score $S(P)$ is defined as

$$S(P) = \sum_{i=1}^{n-1} s_i + \lambda \cdot \sum_{i=1}^{n-2} s_i^r, \quad (3.6)$$

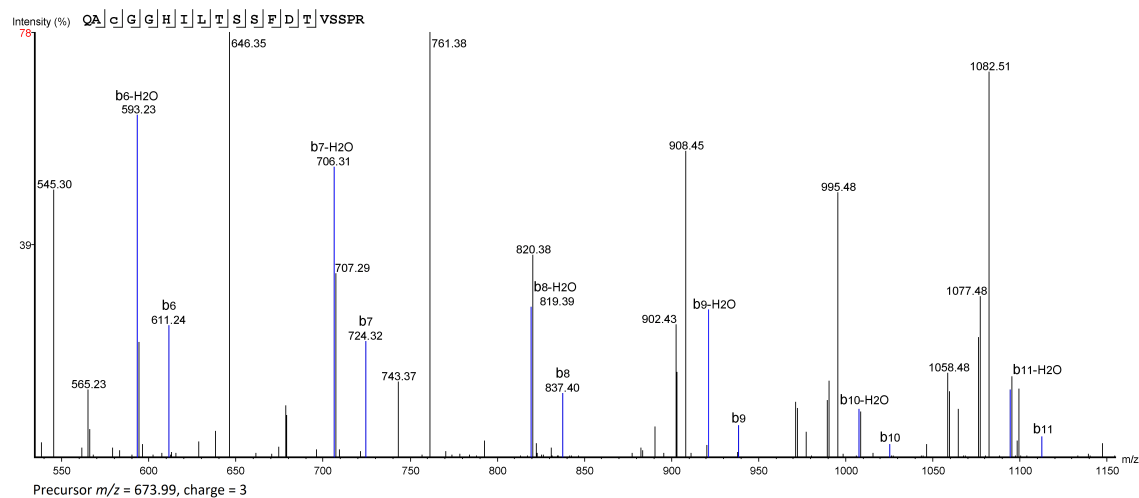


Figure 3.3: An annotated MS/MS spectrum that has stronger b -H₂O ions for many adjacent fragmentation sites, indicating strong correlation between the adjacent fragment ions with the same ion type.

where λ is a positive coefficient that balances the weights of fragmentation site scores and residue scores. In ADEPTS, the optimal λ is determined by performing a grid search on the training data.

3.2 Experimental Results

ADEPTS was applied to two independent MS/MS data sets to evaluate its performance. The first data set (LTQ-Orbitrap data set) was obtained from a Thermo Fisher LTQ Orbitrap XL with ETD mass spectrometer (Thermo Fisher ScientificTM, Bremen, Germany), and the second data set (Iontrap data set) was generated by an ion trap mass spectrometer with implemented ETD module (Model HCTultra PTM discovery system, Bruker Daltonik GmbH, Bremen, Germany). The second data set was previously used by Bertsch *et al.* [16] to evaluate the performance of their CompNovo software.

3.2.1 LTQ-Orbitrap Data Set

In this data set, the CID and ETD data were generated from two separate runs, respectively. Therefore, CID/ETD spectrum pairs needed to be found before using ADEPTS for the *de novo* sequencing.

In our experiment, we used PEAKS to find true spectrum pairs. The database search module of PEAKS was applied separately to two types of spectra for the identification of real peptide sequences, searching in the UniProt database [9]. A pair of spectra from the two runs are considered as from the same peptide if: (1) the two peptide sequences identified for both the CID and ETD spectra are the same; (2) the PEAKS database search confidence score ($-10 \lg P$) is at least 60% on both spectra; and (3) the two spectra have the same charge state and similar retention time (subject to a fluctuation of at most ± 10 minutes). The fairly large LC retention time fluctuation (± 10 minutes) is chosen since the same peptide in two runs is often eluted at different retention time. The first two conditions ensured that the selected CID and ETD spectra in a spectrum pairs are indeed from the same peptide with high confidence. These selected spectrum pairs were regarded as the true spectrum pairs with correct peptide sequences and used for the performance comparison among several software tools later.

317 CID/ETD spectrum pairs with unique peptide sequences were obtained according to the above criteria from the LTQ-Orbitrap data set. 148 of them were randomly chosen as training data, and the remaining 169 were used as testing data. The uniqueness of the peptides guaranteed that no peptide was in both training and testing data. There were 2,291 amino acid residues in the training data and 2,648 residues in the testing data.

A fatal issue prevented the processing of these spectrum pairs by CompNovo. Thus, ADEPTS was only compared with two state-of-the-art software tools, PepNovo and PEAKS, on this data set. The major shortcoming of these two software tools is that only one type of MS/MS spectrum is used to identify the peptide sequence. PEAKS was used to analyze CID and ETD data in two different runs. PepNovo (release 20091029) was only applied to the CID data for its lack of parameters for ETD data analysis.

The precursor mass error tolerance and fragment error tolerance were 0.1 Da and 0.5 Da, respectively. The large fragment error tolerance is due to the measurement

Table 3.2: Comparison between identifications of spectra in the LTQ-Orbitrap testing data set. The first four rows are the percentage of the peptide sequences with at most 0, 1, 2, and 3 incorrect residues in the *de novo* sequencing results. The last row is the percentage of the total correct residues.

	PEAKS (CID)	PEAKS (ETD)	PepNovo (CID)	ADEPTS (CID+ETD)	TOTAL
Correct peptides	12 (7.1%)	10 (5.9%)	8 (4.7%)	32 (18.9%)	169
≤ 1 incorrect residue	12 (7.1%)	13 (7.7%)	11 (6.5%)	33 (19.5%)	169
≤ 2 incorrect residues	25 (14.8%)	25 (14.8%)	27 (16.0%)	66 (39.1%)	169
≤ 3 incorrect residues	28 (16.6%)	34 (20.1%)	33 (19.5%)	74 (43.8%)	169
Total correct residues	945 (35.7%)	1,157 (43.7%)	972 (36.7%)	1,580 (59.7%)	2,648

of fragment ions using ion trap, which provides MS/MS spectra with relatively low resolution as shown in Table 2.1. Table 3.2 illustrates that ADEPTS outperforms both PEAKS and PepNovo on the number of correctly identified peptide sequences and amino acid residues.

3.2.2 Iontrap Data Set

The Iontrap data set was previously published with PRIDE [73, 72] by Bertsch *et al.* to evaluate the performance of their CompNovo software [16]. It contained 156 CID/ETD spectrum pairs as training data and 2,405 pairs ¹ as testing data. The total number of amino acid residues in the training and testing peptides were 1,906 and 32,186, respectively. The fragmentation pattern in this data set was quite different from the one in the LTQ-Orbitrap data set. This difference was expected since different types of mass spectrometers from different manufactures were used to collect the MS/MS data.

¹The testing data set originally contained 2,406 spectrum pairs as described in the paper of CompNovo. One pair was removed in our experiment due to the different precursor m/z values.

Table 3.3: Comparison of identifications on the spectra in the Iontrap testing data. The first four rows are the percentage of the peptides with at most 0, 1, 2, and 3 incorrect residues in the *de novo* sequencing results. The last row is the percentage of the total correct residues.

	PEAKS (CID)	PEAKS (ETD)	PepNovo (CID)	CompNovo (CID/ETD)	ADEPTS (CID/ETD)	TOTAL
Correct peptides	286 (11.9%)	57 (2.4%)	63 (2.6%)	676 (28.1%)	820 (34.1%)	2,405
≤ 1 incorrect residue	301 (12.5%)	74 (3.1%)	99 (4.1%)	697 (29.0%)	844 (35.1%)	2,405
≤ 2 incorrect residues	779 (32.4%)	216 (9.0%)	404 (16.8%)	1,243 (51.7%)	1,350 (56.1%)	2,405
≤ 3 incorrect residues	971 (40.4%)	362 (15.1%)	664 (27.6%)	1,445 (60.1%)	1,528 (63.5%)	2,405
Total correct residues	20,069 (62.4%)	13,767 (42.8%)	19,175 (59.6%)	23,721 (73.7%)	24,378 (75.7%)	32,186

We applied ADEPTS and other *de novo* sequencing software tools on the Iontrap data set. The precursor and fragment mass error tolerance were set as 1.5 Da and 0.4 Da, respectively. We used the same tolerance values with the ones set in the paper of CompNovo because of another failure of CompNovo on its own data set.² The *de novo* sequencing results published by Bertsch *et al.* in their paper were used in the subsequent comparison. Table 3.3 summarizes the results of the comparison. ADEPTS not only significantly beats the *de novo* sequencing accuracy of PEAKS and PepNovo, but also performs noticeably better than CompNovo, which is specially designed to do *de novo* sequencing using CID/ETD spectrum pairs.

Figure 3.4(a) compares the identification rates of different software as the function of the number of allowed incorrect residues in each peptide. Figure 3.4(b) compares the identification rates of different software as the function of the length of the longest consecutively correct subsequence. Given a number of allowed incorrect residues or a number of correct consecutive residues, a higher identification rate indicates better performance achieved by a software tool. Clearly, as shown in Figure 3.4, software tools that combine CID/ETD spectra pair (ADEPTS and CompNovo) perform signif-

²The running issue of CompNovo on its own data set was confirmed by its author after the consultation.

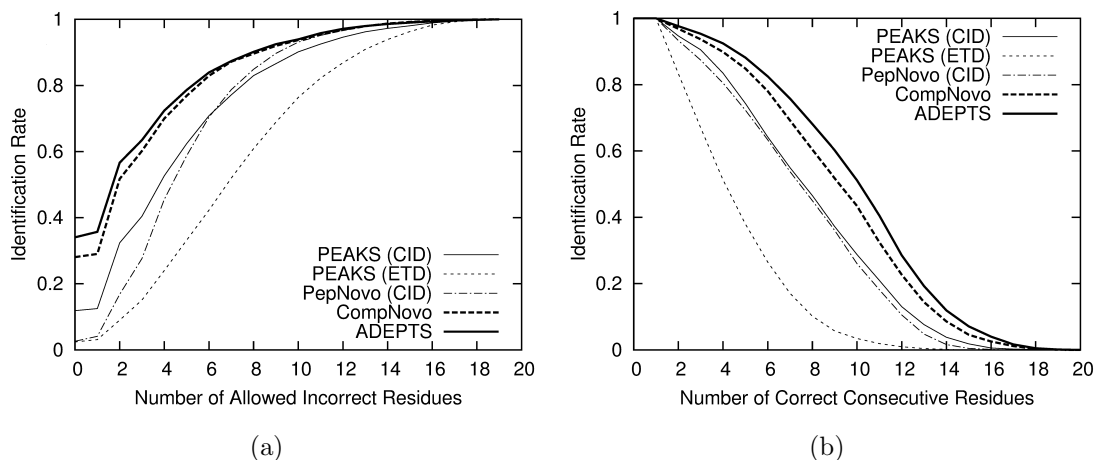


Figure 3.4: (a) Identification rates of different software as the function of the number of allowed incorrect residues for the testing data. (b) Identification rates of different software as the function of the number of allowed incorrect residues for the testing data.

icantly better than the ones that use only one type of spectra (PEAKS and PepNovo). Furthermore, it is also illustrated that the performance of ADEPTS is noticeably better than CompNovo.

3.3 Discussion

It is possible that any *de novo* sequencing tool that generates multiple peptide candidates can be adapted for candidate generation. We used PEAKS because it is regarded as the most superior general *de novo* sequencing software with respect to accuracy and efficiency [120], and it also supports *de novo* sequencing for both CID and ETD. In addition, it has the capability of generating as many as 1,000 candidates for each spectrum. This large number of the peptide candidate increases the probability that the correct peptide is included and finally reported by ADEPTS, since ADEPTS itself does not revise any peptide candidate. As shown in Section 3.2, ADEPTS significantly outperforms PEAKS on both CID and ETD data interpretation. This indicates the importance of the scoring function specially designed for CID/ETD spectrum pairs. In addition, the peptide candidates are currently generated using CID and ETD separately. The performance might be further improved if the candidates were generated from the spectrum pair, rather than only one spectrum.

There are correlations between different ion types at the same fragmentation sites and between two adjacent fragment ions with the same type. To better take account of these correlations in a scoring function, the Bayesian network model used by Datta *et al.* or PepNovo is more appropriate. However, this requires to discretize peak intensity (or significance value), resulting in information loss and thus an apparent decrease of result accuracy. In ADEPTS, instead of constructing a model to count in these correlations directly, an SVR model is used to balance the weight of each type of ions at the same fragmentation site, and the residue score defined in Section 3.1.4 is used to reflect the correlations between adjacent fragment ions. Designing a model that has the power of a Bayesian network but lacks the weakness induced by discretization is still an open problem.

According to the guidance of correctly using LIBSVM given by Hsu *et al.* [69], both the radial basis function (RBF) kernel and the linear kernel were tried in ADEPTS. The parameters were optimized for both kernels and they achieved very similar performance. We selected the linear kernel in our scoring function for its simplicity and efficiency, conforming to Occam’s razor.

Instead of using the likelihood scores to form the vector for SVR score calculation, we also tried to use the peak significance values directly, but the accuracy was remarkably reduced. This also indicates that the conversion from significance values to likelihood scores is necessary.

Our method requires a CID/ETD spectrum pair from the same peptide. This can be obtained either by (1) two separate LC-MS/MS runs of the same protein digest, or by (2) programming the mass spectrometer to fragment the same precursor ion in two consecutive scans using CID and ETD, respectively. For the first setting, some treatments on the data are needed to obtain spectrum pairs. In general, data-dependent acquisition (DDA) mode used in an MS/MS spectrometer for data collection favors not fragmenting one peptide repeatedly, thus spectrum pairs can be readily selected by a trivial procedure: given two spectra from CID and ETD respectively, they are regarded to form a spectrum pair if they have similar precursor m/z values (the difference is less than or equal to a given threshold, *e.g.*, 10 ppm for Orbitrap data) and similar retention time. Moreover, the pairing of peptide features from two LC-MS experiments is well studied in the label-free quantification method [139] and can be readily facilitate the spectrum pair finding. For un-paired spectra, we can still use the traditional *de novo* sequencing methods for the analysis.

Chapter 4

De Novo Sequencing with Many PTMs

The identification of post-translational modifications (PTMs) is of critical importance in a study of protein functions. Novel proteins can be identified using *de novo* sequencing approaches, but identifying PTMs on the proteins is a nontrivial challenge.

Most *de novo* sequencing software tools allow users to specify only a few PTMs. This limitation makes the peptides with unspecified PTMs become unidentifiable. As pointed out by Duncan *et al.* [39], peptides with unspecified PTMs may be especially interesting, but unfortunately they are discarded due to limitations of existing software.

A natural solution is to specify all the PTMs that possibly exist in a sample. This approach only linearly increases the time complexity of commonly used *de novo* sequencing algorithms. However, the consideration of a large number of PTM types typically has a significantly negative impact on the result accuracy. *De novo* sequencing algorithms, such as PEAKS and PepNovo, tend to output peptide sequences with many PTMs in such case, though most peptides identified in a proteomic study have very few, if any, PTM occurrences in each peptide.

It is different between the number of PTM occurrences per peptide and the number of PTM types specified to a *de novo* sequencing algorithm. Researchers usually do not know all the PTM types in a peptide and thus have to specify a large number of PTM types. However, only a limited number of PTMs can occur in a peptide. Therefore, we develop a specialized *de novo* sequencing algorithm, DeNovoPTM, which allows

the consideration of many PTM types, while limiting the number of PTM occurrences in each peptide.

4.1 Problem Formulation

Let S be a given MS/MS spectrum. The total residue mass, M , of the peptide can be derived from the precursor m/z and the charge of the spectrum. The spectrum is usually provided with a peak list. Each peak (m_i, h_i) potentially corresponds to a fragment ion, where m_i represents the m/z value of the peaks and h_i is its abundance.

Let $\Sigma = \{r_1, \dots, r_{|\Sigma|}\}$ be the alphabet of amino acid residues. Besides 20 *unmodified residues*, this alphabet also includes *modified residues* that represent residues with PTMs. Each residue r_i has mass value $m(r_i)$. A peptide P is an sequence of amino acid residues over alphabet Σ and $m(P)$ denotes the total mass of residues in P . A scoring function $F(P, S)$ is required to evaluate the similarity between a peptide sequence P and a given MS/MS spectrum S , where a higher score means a higher probability that the given spectrum is generated from the peptide.

The *de novo* sequencing approach involves constructing a peptide sequence P from a spectrum S over the alphabet Σ , such that (1) the total residue mass is M and (2) the similarity score $F(P, S)$ is maximized. It is further required that the number of modified residues in the computed peptide sequence P is upper-bounded by a given number k . Formally, the *de novo* sequencing problem with a limited number of PTMs per peptide (DeNovo-LPTM), given a scoring function F , is defined as follows:

DeNovo-LPTM

Instance: An MS/MS spectrum S , the precursor mass M , a residue alphabet Σ , and a maximum number of PTMs per peptide k .

Objective: A peptide sequence P , over alphabet Σ , that satisfies: (1) the total residue mass $m(P)$ is equal to M (within a specified error tolerance), (2) the number of modified residues in P is no more than k , and (3) the similarity score $F(P, S)$ is maximized.

4.2 Methods

In this section, the scoring function $F(P, S)$ is introduced, and the algorithm for the DeNovo-LPTM problem is then proposed.

4.2.1 Scoring Function

The scoring function we use for *de novo* sequencing here is similar to the one used by ADEPTS, discussed in previous chapter, where a CID/ETD spectrum pair is used to identify the target peptide. Here, the scoring function is slightly modified and applied to CID spectra only.

Peaks in a spectrum are first assigned non-negative significance values using Equation 3.1. A match between a peak and a theoretical fragment ion of a proposed peptide sequence contributes to the correctness of the corresponding peptide candidate. This contribution is measured by a likelihood score according to its significance value and the type of the matched fragment ion. A likelihood score vector for a fragmentation site is generated and fed to a pre-trained SVR model to calculate the final score for the corresponding fragmentation site. Thus, for each mass value m , a score $f(m)$ can be computed to represent the likelihood that the peptide has a prefix with the total residue mass m . The fragmentation score of the peptide P is the sum of all scores at all prefix mass values of the peptide.

In addition, our scoring function includes a penalty for each modified residue in a peptide. If an MS/MS spectrum can be explained by two peptides, one has a PTM but the other does not, with the same scores, we prefer the one without the PTM [60, 127]. The value of this penalty should vary according to the frequency of the specific PTM type being observed in a proteomic experiment. While our algorithm can work for any user-defined penalties, the following configuration performs well in our experiment. All PTMs in the Unimod database [31] were empirically classified into four classes: common, less common, rare and very rare, and were assigned with penalties of -0.15, -0.3, -0.45 and -0.6, respectively.¹ For the sake of presentation, we denote the penalty for a residue r by $g(r)$. If r is an unmodified residue, $g(r) = 0$; otherwise, $g(r) < 0$. The score $F(P, S)$ is defined as the sum of the fragmentation score and the PTM penalties.

¹The classification and the four penalty values are configurable in the software.

4.2.2 DeNovoPTM Algorithm

For a deconvoluted and de-isotoped spectrum, the m/z value of each peak is converted to its nominal mass by multiplying 0.9995 and rounding to the nearest integer. The constant 0.9995 is the average ratio between the nominal and the accurate mass of the 20 basic amino acid residues [14, 67]. If more than one peak exist in an integer bin, those peaks are treated as one and their corresponding intensity values are added together as the new intensity. The total residue mass M is rounded into an integer in the same way. Such integralization is to facilitate the use of a dynamic programming algorithm for the *de novo* sequencing.

The score $F(P, S)$ can be similarly defined on a partial sequence p of which the total residue mass is less than M . For each mass $m \leq M$, there is an optimal partial sequences p with mass m such that the score $F(p, S)$ is maximized. If there are multiple partial sequences with mass m that maximize the score, any of them satisfies our purpose.

Our algorithm maintains a $(k + 1)$ by M matrix DP , in which $DP(i, m)$ denotes the optimal score that can be achieved by a partial sequence with mass m and i PTMs. The optimal partial sequence p for $DP(i, m)$ must be of the form $p'r$, consisting of a prefix sequence p' and a residue $r \in \Sigma$. It is clear that $F(p, S) = F(p', S) + f(m) + g(r)$. Furthermore, p' must also be the optimal partial sequence with its mass $m' = m - m(r)$; otherwise, we can replace p' by a better partial sequence to improve the score of p , and this is a contradiction to the optimality of p .

Let $\Sigma_0 \subset \Sigma$ denote the set of unmodified residues in the alphabet, and $\Sigma_1 \subset \Sigma$ be the set of modified residues. The following recurrence relation is explicit because of the above discussion:

$$DP(i, m) = f(m) + \max \begin{cases} \max_{r \in \Sigma_0} DP(i, m - m(r)) \\ \max_{r \in \Sigma_1} DP(i - 1, m - m(r)) + g(r) \end{cases} \quad (4.1)$$

The dynamic programming algorithm for the DeNovo-LPTM problem, DeNovoPTM, is shown in Algorithm 1. The time complexity of the DeNovoPTM algorithm is $O(kM|\Sigma|)$.

It is known that a *de novo* sequencing algorithm has a tendency to output a peptide that matches the high peaks in the spectrum with multiple fragment ions [98, 34]. To

Algorithm 1 DeNovoPTM algorithm to solve the DeNovo-LPTM problem.

Require: An MS/MS spectrum S , the precursor mass M , an unmodified residue set Σ_0 , a modified residue set Σ_1 , and a maximum number of PTMs per peptide k .

```

1: function DENOVOPTM( $S, M, \Sigma_0, \Sigma_1, k$ )
2:    $DP[i, m] \leftarrow 0$  for  $0 \leq i \leq k + 1, 0 \leq m \leq M$ 
3:   for  $i \leftarrow 0$  to  $k + 1$  do
4:     for  $m \leftarrow 1$  to  $M$  do
5:        $DP(i, m) = f(m) + \max \begin{cases} \max_{r \in \Sigma_0} DP(i, m - m(r)) \\ \max_{r \in \Sigma_1} DP(i - 1, m - m(r)) + g(r) \end{cases}$ 

```

avoid this problem, the peptide reported by our algorithm is checked. If there is one peak in the spectrum matched by both an N-terminal fragment ion and a C-terminal fragment ion, we run the dynamic programming twice, in one run forbidding the peak from being matched by an N-terminal, and in the other forbidding a C-terminal fragment ion. If there are t significant peaks found to be matched by different ion types, the algorithm is run 2^t more times. In practice t is often 0 or a very small integer. A similar strategy was also previously proposed in Mo *et al* [107].

4.3 Experiments and Results

We implemented our DeNovoPTM algorithm for *de novo* sequencing with a limited number of PTMs. The performance of DeNovoPTM was evaluated by comparing with two state-of-the-art *de novo* sequencing software tools, PEAKS (*de novo* sequencing module) [98] and PepNovo (Release 20120423) [45], on two data sets: the ISB data set and the PepSplice data set.

4.3.1 Performance Evaluation on the ISB Data Set

The MS/MS spectra came from the ISB (Institute for System Biology) standard protein mixture data set [82], which is a standard protein data set for testing peptide identification software tools. In our experiment, data from two LC-MS/MS runs for the analysis of mixture 3 and 7 were selected. Mixture 3 was analyzed by an Agilent

1100 system, while mixture 7 was analyzed by a Thermo Scientific Orbitrap MS/MS spectrometer.

To determine the control set, we used the database search module of PEAKS 6 [166] to identify the peptide-spectrum matches (PSMs) with high confidence. Four PTMs were used in the database search: carbamidomethylation on Cys, oxidation on Met, deamidation on Asn and Gln, and phosphorylation on Ser, Thr and Tyr. All the spectra were searched against the 18 standard proteins and the possible contaminant proteins given by ISB. The reported modified PSMs are filtered using the following rules: (1) PEAKS $-10 \lg P$ score of a PSM must be greater than 35; (2) a PSM must have at least one cysteine carbamidomethylation or one phosphorylated amino acid; (3) the precursor charge of the spectrum is 2; (4) for the PSMs with same peptides, we keep the one with the highest $-10 \lg P$ score. After this strict filtration, 85 modified PSMs with high confidence were obtained. These PSMs contained 1,094 residues, with 120 modified residues.

We selected 71 frequently observed PTMs, using as common PTMs in PEAKS, as possible PTMs in our experiment. All these PTMs were set as variable PTMs. The number of PTM sites per peptide was limited to two in both PEAKS and DeNovoPTM. PepNovo does not support such a limitation of PTM number, so no limitation was specified.

To study the effect of the number of PTM types on the performance, each software tool was run three times by specifying the four real PTMs mentioned above, the 38 PTMs in Table 4.1, and all 71 frequently observed PTMs in PEAKS, respectively.

Table 4.2 listed the comparison of three *de novo* sequencing tools in terms of the number of residues and modified residues correctly reported by each tool. As a *de novo* sequencing algorithm cannot distinguish two residues (or modified residues) with the same mass, a reported residue (or modified residue) is regarded correct if its mass is equal or similar (with mass error up to 0.1 Da) to the real possibly modified residue at the same position of the peptide. Table 4.2 shows that DeNovoPTM outperforms PEAKS and PepNovo on the number of correct PTM sites identified in all three experiments. In terms of the number of correctly identified amino acids, PEAKS performs the best when only four PTM types are used, while DeNovoPTM gets the first rank in the experiments with 38 and 71 PTMs being specified. This indicates the advantage of our algorithm when many PTM types are specified. Moreover, we notice that the performance of PEAKS and PepNovo degrades significantly when the

4.3. EXPERIMENTS AND RESULTS

Table 4.1: Thirty-eight PTMs used to evaluate the performance of three *de novo* sequencing software tool.

Index	Mass	Residues	PTM Name
1	57.02	C	Iodoacetamide derivative (C)
2	42.01	K, X@N-term	Acetylation (K, X@N-term)
3	0.98	NQ	Deamidation (NQ)
4	79.97	STY	Phosphorylation (STY)
5	14.02	DE, X@C-term	Methylation (DE, X@C-term)
6	15.99	M	Oxidation (M)
7	79.96	Y	O-Sulfonation (YTS)
8	42.05	RK	tri-Methylation
9	-0.98	X@C-term	Amidation
10	43.01	K, X@N-term	Carbamylation (K, X@N-term)
11	43.99	EDKW	Carboxylation
12	14.02	RK	Methylation (RK)
13	-29.99	M@C-term	Homoserine
14	-48.00	M@C-term	Homoserine lactone
15	99.07	C	N-isopropylcarboxamidomethyl
16	-18.01	C@N-term	Dehydration (C@N-term)
17	71.04	C	Acrylamide adduct (C)
18	39.99	C@N-term	S-carbamoylmethylcysteine cyclization
19	-18.01	E@N-term	Pyro-glu from E
20	-17.03	Q@N-term	Pyro-glu from Q
21	21.98	DE, X@C-term	Sodium adduct
22	105.06	C	S-pyridylethylation
23	15.99	WH	Oxidation or Hydroxylation (WH)
24	45.99	C	Beta-methylthiolation (C)
25	42.02	K	Guanidination
26	27.99	X@N-term	Formylation (X@N-term)
27	44.03	C	Ethanolation (C)
28	-17.03	C@N-term	Loss of ammonia (C@N-term)
29	31.99	M	dihydroxy
30	162.05	T	Hexose (T)
31	203.08	N	N-Acetylhexosamine (N)
32	210.20	CK, G@N-term	Myristoylation
33	226.08	K, X@N-term	Biotinylation
34	42.01	TSCYH	Acetylation
35	227.13	C	Applied Biosystems cleavable ICAT(TM) light
36	236.16	C	Applied Biosystems cleavable ICAT(TM) heavy
37	0.98	R	Deamidation (R)
38	27.99	TKS	Formylation (TKS)

Table 4.2: Comparison between the performances of three *de novo* sequencing software tools on the ISB data set. Each software tool was run three times with 4 PTMs, 38 PTMs, and 71 PTMs being specified, respectively. This table listed the numbers of PTMs that were identified correctly on both PTM types and positions, as well as the numbers of correctly identified residues.

Software	PEAKS			PepNovo			DeNovoPTM			Real
Number of specified PTMs	4	38	71	4	38	71	4	38	71	
Number of correct PTMs	78 (65%)	41 (34%)	24 (20%)	64 (53%)	50 (42%)	41 (34%)	79 (66%)	65 (54%)	61 (51%)	120
Number of correct residues	725 (66%)	569 (50%)	486 (44%)	675 (62%)	595 (54%)	562 (51%)	691 (63%)	641 (59%)	653 (60%)	1,094

number of PTM types increases, whereas the performance of DeNovoPTM degrades slowly.

4.3.2 Performance Evaluation on the PepSplice Data Set

To further evaluate performance, we applied DeNovoPTM to another data set obtained from an ion trap MS/MS spectrometer. This data set was previously used by Roos *et al.* in their PepSplice paper [127]. We downloaded the PepSplice data set after the development of the DeNovoPTM software and the selection of the parameters; therefore, the test on this data set can be regarded as a blind test. The MS/MS data were searched against the UniProt database [9] using the PTM search module of PEAKS [60]. We used 0.5 Da as the precursor and fragment error tolerance, considering the low precision of ion trap instruments. Among the 195,314 MS/MS spectra, PEAKS identified 2,020 unique modified peptides with high confidence ($-10 \lg P \geq 35$), including 12 types of PTMs. We used these 2,020 modified PSMs as the control set in the subsequent evaluation.

The previously mentioned 71 PTMs were used as variable PTMs for three *de novo* sequencing software tools: PEAKS, PepNovo, and DeNovoPTM. Two variable

Table 4.3: Comparison between the performances of three *de novo* sequencing software tools on the PepSplice data set. Each software tool was run three times with 4 PTMs, 38 PTMs, and 71 PTMs being specified, respectively. This table listed the numbers of PTMs that were identified correctly on both PTM types, as well as the numbers of correctly identified residues.

Software	PEAKS			PepNovo			DeNovoPTM			Real
Number of specified PTMs	4	38	71	4	38	71	4	38	71	
Number of correct PTMs	425 (16%)	245 (9%)	165 (6%)	433 (16%)	336 (13%)	267 (10%)	415 (16%)	390 (15%)	366 (14%)	2,672
Number of correct residues	13,652 (44%)	9,479 (30%)	7,544 (24%)	12,821 (41%)	10,601 (34%)	9,835 (31%)	12,167 (39%)	11,796 (38%)	11,785 (38%)	31,310

PTMs were allowed per peptide in both PEAKS and DeNovoPTM. The precursor and fragment error tolerance values of all three software tools were set as 0.5 Da. Table 4.3 lists the performance comparison of these three software tools on this ion trap data set. This comparison, similar to the previous one, shows that the performance of our algorithm is the best compared with other two tools when many variable PTMs are involved. Furthermore, it also shows that the performance of DeNovoPTM degrades more slowly when the number of involved PTMs is increased.

4.4 Discussion

In this chapter, an efficient dynamic programming algorithm, DeNovoPTM, was proposed and implemented for *de novo* sequencing. DeNovoPTM can be regarded as a specialized *de novo* sequencing algorithm for a particular application. The experimental results show that our algorithm outperforms two state-of-the-art *de novo* sequencing algorithms, PEAKS and PepNovo, when the number of possible PTM types is large. Particularly, the better performance comes from the ability to limit the number of PTM sites per peptide in our algorithm.

In our study, the PTM types are unknown in advance, and researchers have to

turn on many possible PTM types. This increases the solution space of the *de novo* sequencing problem and leads to a higher chance of false positives and worse performance of general tools. However, most peptides only contain a limited number of PTM sites per peptide. Therefore, by using a specifically designed algorithm to limit the number of PTM sites per peptide, our algorithm efficiently reduces the solution space. This contributes to an improvement of *de novo* sequencing accuracy as shown in the experimental section.

In some other situations, the researchers actually know additional information and then special algorithms can be designed to achieve better performance by utilizing more information than the general tools. For example, Bahtia *et al.* [17] added previously known peptide patterns to help improving the accuracy of *de novo* sequencing. However, in many cases, the performance improvement is mostly due to the utilization of the additional information or the correct handling of the lack of information. It does not necessarily mean that the special tools will, or need to outperform the general tools for all situations, and such a phenomenon has been shown in the two performance comparisons.

The software implementation of our algorithm rounded all the mass values to the nominal (integer) mass values. This will lose some information when a high resolution mass spectrometer is used. However, the algorithm can be easily adjusted to utilize the high mass accuracy through multiplying each mass value by a large integer before the rounding.

Although there are more than 600 PTMs in the Unimod database, only 71 PTMs used in our experiments are listed in PEAKS as the most commonly observed ones. In theory all the PTMs in the Unimod database can be specified to DeNovoPTM; however, this is not recommended because of the serious reduction of *de novo* sequencing accuracy. For the identification of rarely observed PTMs, we cannot solely rely on a single MS/MS spectrum obtained in a high-throughput proteomic experiment.

Besides the *de novo* sequencing application, another possible application of DeNovoPTM algorithm is to provide a short list of the most likely PTMs from a large number of PTMs provided by users. The MS/MS data can be reanalyzed using a traditional software tool by only considering the short list of PTMs.

Chapter 5

Database Search for Modified Peptides Without Specifying PTMs

During the past decades, many database search software tools have been developed for peptide identification from MS/MS data [28, 41, 52, 118, 166]. However, these software tools provide limited support to modified peptide identification using a straightforward procedure proposed by Yates *et al.* [164]: users specify the PTMs that possibly exist in the sample. These search tools are regarded as *conventional* (or *traditional*) database search engines.

PTMs specified by users are often categorized into fixed and variable PTMs. If a PTM is specified as fixed, every occurrence of the residue will be replaced with the modified residue and the consideration of these fixed PTMs will not affect the software's running time. In contrast, the consideration of variable PTMs dramatically increases the workload of computing. In particular, many variable PTMs can modify multiple amino acid residues in a peptide, easily causing an exponential growth of search space.

The search space growth increases not only the running time, but also the potential false discoveries to an unacceptable level. Therefore, when a conventional database search engine is used for peptide identification, only a few variable PTMs can be practically specified, while peptides with unspecified PTMs will not be reported. Some researchers regard such limitation on conventional database search engines as

one of the major factors that contribute to the current low identification rate of MS/MS spectra [146] and the low characterization rate of modified peptides [39].

There exist software tools that have been developed for the identification of unspecified PTMs. Many sequence tag-based tools, including the first tag-based database search algorithm by Mann *et al.* [101], GutenTag [144], OpenSea [133] and SPIDER [61], can be used to identify modified or mutated peptides from a protein database. In these approaches, peptide sequence tags are generated from a spectrum using *de novo* sequencing and then searched for the approximate matches in a protein database. The differences between the tag and a matched peptide from the database can be explained by either mutations or PTMs. InsPecT [146], MODⁱ [79] and ByOnic [14] employ hybrid search approaches: InsPecT speeds up the database search through using partial *de novo* sequencing tags to select peptide candidates, whereas the actual comparison between the spectrum and the peptide sequence is achieved by a 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. MODⁱ applies a straightforward algorithm to search for modified peptides in a protein database with at most 20 proteins. The small number of proteins is insufficient for the study of complex protein mixtures. ByOnic uses “lookup peaks” to extract peptide candidates from a protein database. Commercial software tools such as the Paragon algorithm [135] and Mascot (Error Tolerant Search Mode) [30] take a large number of PTMs in consideration during the search. To avoid the combinatorial explosion of the search space, Paragon uses *de novo* sequencing tags to locate “hot” areas in the protein database, where PTMs are intensively checked, while Mascot only allows one type of PTM per peptide.

Several software tools have recently benefited from the discovery that many modified peptides have their *unmodified forms (base forms)* co-existing in the data. For example, MS-Alignment [151] uses a dynamic programming algorithm to compare a pair of spectra that are possibly generated by the modified and the base forms of a peptide respectively. ModifiComb [130] uses the same principle except that a pair of spectra is compared with each other only if one spectrum is identified as an unmodified peptide. The retention time difference between the base and the modified forms is also considered in ModifiComb. Another study [10] has further extended this principle to form a spectral network using differently modified forms of the same peptide.

In this chapter we present a novel software tool, PeaksPTM, for modified peptide identification without specifying PTMs. The first improvement is a default setting by which the software considers all PTMs included in the Unimod database as variable PTMs. We then add several search strategies to reduce the search space. The scoring function in PeaksPTM uses the co-existence of modified and base forms of the same peptide in a more effective way than either MS-Alignment or ModifiComb. Our experiments show that PeaksPTM performs better on modified peptide identification than four other state-of-the-art software tools.

5.1 Methods

PeaksPTM is designed for modified peptide identification from spectral data generated by typical LC-MS/MS experiments. It makes use of the high mass accuracy of precursor ions from survey scans, generated from a high-resolution mass spectrometer. The MS/MS data can be measured with a low-resolution mass analyzer.

PeaksPTM adopts a two-pass database search strategy. In the first pass, a traditional database search module identifies a list of possible proteins, with only a few commonly observed PTMs specified. In the second pass, modified peptides from this short list of proteins are checked for each spectrum with the consideration of all the PTMs in the Unimod database. The computational analysis consists of four major steps:

1. Protein identification. The MS/MS spectra are searched against a protein database by PEAKS database search module for the identification of a short list of protein candidates. This will filter out most impossible proteins to significantly decrease the search space for the next step.
2. Single-PTM peptide candidate search. Protein candidates are digested *in silico* into a set of peptide candidates. For each spectrum, an exhaustive search is performed to find all corresponding peptide candidates with a limitation of at most one variable PTM per peptide. This one-PTM-per-peptide limitation avoids the exponential growth of the search space.
3. Peptide candidate rescoring. The peptide candidates for a spectrum are rescored by combining three features: the *LDF score* calculated by PEAKS, the *peptide pair* and *PTM rareness*.

- PEAKS LDF score. This score uses a linear discriminant function (LDF) that involves three features: the PEAKS PSM score [97], the peptide length and the average score of the 512 best PSM scores for the spectrum.
 - Peptide pair. This feature examines a modified peptide candidate to determine if its base form can be independently identified from another spectrum. The co-identification of both modified and base forms of the same peptide increases the identification confidence.
 - PTM rareness. A modified peptide with a rare PTM has to obtain a higher PEAKS LDF score to receive the same level of confidence as a peptide modified by a common PTM. This feature adjusts the score of a modified peptide candidate according to the commonality of the PTM.
4. Multi-PTM peptide search. Common PTMs identified in single-PTM peptide search are used to search for modified peptides with two or more PTMs.

PeaksPTM also controls result quality by a modified target-decoy approach, following the proposal designed for two-pass database search approaches by Bern *et al.* [15]. Moreover, we also propose a straightforward and effective strategy to combine the results from multiple search engines to further improve the identification rate.

The details of the analytical steps, the features for rescoring, the quality control of the result, and the consensus strategy are discussed in the following sections.

5.1.1 Protein Identification

A protein sequence is digested into a set of peptides. Only a partial set of the peptides is fragmented and even less are identified by database search in an MS/MS experiment. However, a protein from a database can still be identified even if only a few peptides of the protein are identified. A short list of protein candidates can thus be obtained by the base peptides and the modified peptides with specified PTMs using PEAKS.

The database search is performed on a pre-constructed target-decoy database to estimate the false discovery rate (FDR) of the identification result. The decoy protein database is generated by shuffling each protein sequence in the target protein

database. To shuffle a protein sequence, the amino acid residues between every two adjacent digestion sites are randomly permuted, while the residue at the digestion site is unchanged. If a shuffled peptide occurred in a target protein, it is removed from the decoy database. The first-round database search on the target-decoy database identifies a short list of protein candidates, including both target and decoy proteins. A *reduced protein database* is constructed using these proteins for the subsequent modified peptide search.

5.1.2 Single-PTM Peptide Candidate Search

Each protein in the reduced protein database is digested *in silico* to peptides, which are regarded as base-form peptides. Single-PTM peptides are then generated by replacing amino acid residues with modified ones. Suppose each amino acid residue has m different PTMs on average in the Unimod database, then for a peptide with length k , mk single-PTM peptides will be generated. This is only a linear growth on the number of peptide candidates. Thus, a brute-force algorithm is used instead of the sophisticated dynamic programming algorithm of InsPecT.

For each spectrum, a peptide candidate, either in base form or modified, is selected for PEAKS PSM score calculation if the difference between the precursor mass of the spectrum and that of the peptide candidate is within a specified mass error tolerance. The top 512 peptide candidates according to the PEAKS PSM scores are selected and each of them is further evaluated by the PEAKS LDF scoring function. The peptide with the top LDF score is kept for each spectrum as its peptide candidate. This peptide candidate for a spectrum can be either a base-form or a single-PTM peptide.

5.1.3 Modified Peptide Rescoring

The peptide candidate of each spectrum is rescored since LDF score is originally optimized for the identification of base-form peptides. To measure the match between a modified peptide and a spectrum, the influence of the PTM in the candidate needs to be considered. LDF scores help to determine peptide candidates for the spectra, and other two features, peptide pair and PTM rareness, are used to take account of PTMs in the peptide candidate.

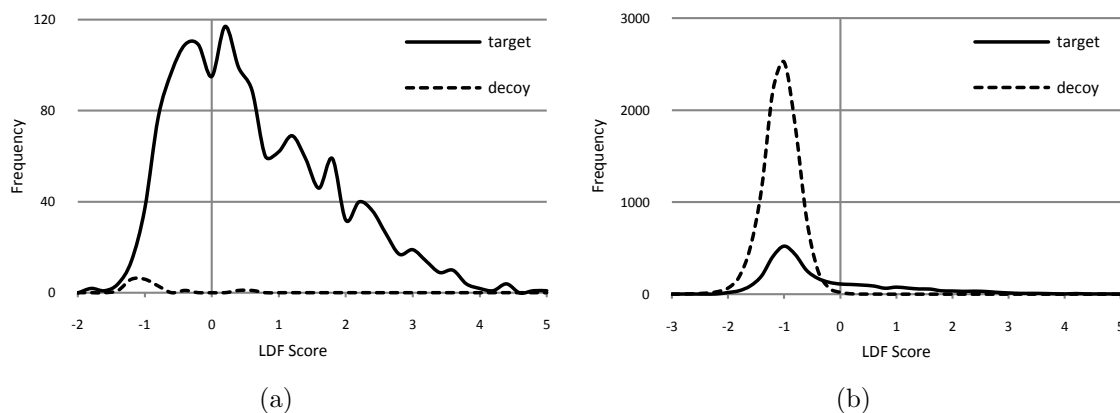


Figure 5.1: The LDF score distributions of single-PTM peptides identified from target and decoy databases, respectively. (a) The distribution with peptide pairs, and (b) without peptide pairs. Modified peptides from the target database tend to have more peptide pairs than those from the decoy database.

Peptide Pairs

Similar to the observations in MS-Alignment [151] and ModifiComb [130], many peptides have spectra in the data set for both their modified and base forms. It is natural to conclude that if both forms of the same peptide are independently identified from different spectra, the identification tends to be correct. This property is illustrated in Figure 5.1 where the peptide pairs found in the target database are significantly more than those found in the decoy database. This discovery is particularly relevant to the modified peptide candidates identified with higher LDF scores and strongly suggests the correctness of the above conclusion.

PeaksPTM uses this peptide pair feature by adding a reward to a modified peptide identification if its base form is independently identified from another spectrum. The reward addition occurs only after the peptide identification. This score adjustment does not change the peptide result but 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 identification of the modified peptide. Compared to previous software, PeaksPTM appears to be less sensitive since some modified peptides may not be identifiable by their spectrum alone. However, the specificity of our method is much improved because it is very rare that two independent identifications constitute both the base and modified forms

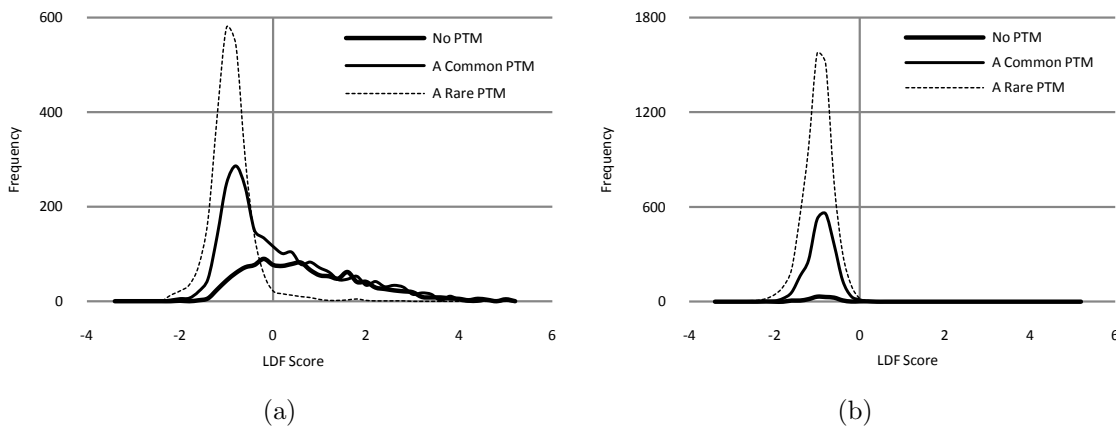


Figure 5.2: The LDF score distributions of the peptide candidates identified with no PTM, a common PTM, and a rare PTM, from (a) the target database and (b) the decoy database.

of the same peptide unless both identifications are correct.

PTM Rareness

A rare PTM in a peptide typically demands a higher LDF score of the peptide to justify its correctness, whereas common PTMs, such as oxidation on Met, are so ubiquitous that their occurrence does not require a higher LDF score than the unmodified peptide. By summarizing the common PTMs reported in previous publications [100, 55], we regard the 29 PTMs in Table 5.1 as common PTMs, and all other PTMs as rare ones.

Figure 5.2 shows the different LDF score distributions of the single-PTM peptide candidates with different PTM types from the target and decoy proteins, respectively. A great distinction is shown on this feature in the target peptides but not in the decoy ones. It suggests a strong correlation between the PTM rareness and the identification correctness.

Since there is no quantitative measurement 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 in a peptide and penalties for both common and rare PTMs are obtained from training. The penalty for a modified peptide, either with single PTM or multiple PTMs, is actually the sum of the PTM penalties.

CHAPTER 5. UNRESTRICTED PTM SEARCH

Table 5.1: The summary of 29 PTMs which are frequently reported in previous research.

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

Weighted Sum Score

Our final score for a modified peptide candidate is a linear combination of four features: the PEAKS LDF score (S_{ldf}), the existence of a peptide pair ($E_{peptide_pair}$), the number of common PTMs (N_{common_ptm}), and the number of rare PTMs (N_{rare_ptm}). More specifically, the scoring function $f(\cdot)$ for calculating the score of a modified peptide candidate P is defined as

$$f(P) = S_{ldf} + c_1 \cdot E_{peptide_pair} - c_2 \cdot N_{common_ptm} - c_3 \cdot N_{rare_ptm} \quad (5.1)$$

where $E_{peptide_pair} = 1$ if there is a peptide pair; otherwise, $E_{peptide_pair} = 0$. The coefficients c_1 , c_2 , and c_3 are obtained by training. This scoring function is also used for rescored modified peptides with multiple PTMs.

The obstacle to determine the coefficients c_i is to find a training data set consisting of a large number of spectra annotated by modified peptides. Manually annotating a large-scale data set is impractical, while simulated data sets used in previous research introduce difficulties to evaluate false negatives [146]. Alternatively, the coefficients can be trained by maximizing the number of identifications at 1% FDR, which is estimated with a target-decoy approach.

5.1.4 Estimation of the False Discovery Rate

The first-round database search on the target-decoy database identifies a short list of protein candidates, including both target and decoy proteins. In general, the decoy proteins in this short list are fewer than the target proteins, and this can result in an underestimation of FDR after the second-round search.

A modified target-decoy strategy, which is specifically designed for two-pass database search approaches, is adopted to avoid the underestimation of FDR [15]. In the first-round search, a target-decoy protein database is searched to determine the possible proteins. The reduced protein database is then constructed using the identified proteins in the first-round search. It contains the target proteins (P_t), the decoy proteins (P_d), and the shuffled proteins generated from P_t . The second-round database search is performed on this reduced protein database to identify modified peptides. This method is only slightly biased against target peptides and the estimated FDR will not be lower than its actual value.

The following method is used to calculate the FDR: suppose there are N_d identifications from the decoy proteins and N_t identifications from the target proteins, the FDR after removing the decoy hits from the results is then calculated as N_d/N_t .

5.2 Experiments and Results

We compared PeaksPTM with Mascot (Mascot 2.3, Error Tolerant Search Mode) [30], Paragon (ProteinPilot software 4.0.8085, Paragon Algorithm: 4.0.0.0, 148083, trial version) [135], and InsPecT (release 20101012) [146] on an MS/MS data set obtained from human heart tissue. PEAKS was applied to generate the short lists of proteins for PeaksPTM and InsPecT. In our experiments, we also compared PeaksPTM with MODⁱ [79].

5.2.1 Data Sets

Two data sets were involved in our experiments:

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 ScientificTM, Bremen, Germany) using a reversed phase analytical column. The data was collected with an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher ScientificTM, Bremen, Germany), consisting of 11,207 survey scans and 15,117 MS/MS spectra.

Yeast: The yeast data set was generated from a fraction of Lys-C digest of a yeast lysate by an LTQ Orbitrap XL mass spectrometer (Thermo Fisher ScientificTM, Bremen, Germany). It contains 5,136 survey scans and 12,366 MS/MS spectra.

5.2.2 Coefficient Determination

The independent yeast data set was used to train the coefficients in Eq. 5.1 for the final score calculation. This was to eliminate the overfitting problem caused by training on the data sets from the same or similar species.

Table 5.2: The numbers of identified peptides with $FDR \leq 1\%$ 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

The performance by such training strategy was verified as shown in Table 5.2. The parameters trained on one data set were used to identify the modified peptides from the other. When the parameters trained on the 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. Conversely, when the parameters trained from the Human-heart data set were tested on the Yeast data set, the number of identifications decreased from 2,447 to 2,410. Using the training and testing data from the same species only produces slightly better results than from different species. This indicates that the overfitting problem in our method is negligible.

5.2.3 Comparison between Multiple Search Engines

PeaksPTM was compared with Mascot, Paragon and InsPecT to evaluate its performance. As Mascot and Paragon have their own first-round search functions, the IPI Human (v3.75) database, concatenated with its shuffled protein sequences, was used as the target-decoy database. The corresponding FDRs were calculated using the standard target-decoy approach [77, 75]. PeaksPTM used the same target-decoy database and found 1,349 target and 773 decoy proteins. A reduced protein database consisting of 3,471 entries was constructed as described above. InsPecT could not finish the whole IPI human database in its blind search mode; therefore, it was applied on a list of 2,030 proteins identified by PEAKS from the target database. This pre-selected protein list was believed to be a superset of the high abundance proteins in the sample. Meanwhile, the decoy protein sequences generated from these 2,030 proteins were also searched to determine the FDR.

For PeaksPTM and Mascot, the precursor and fragment ion error tolerance values were set to 10 ppm and 0.5 Da, respectively. The maximum variable PTM number per peptide was set to 1 in PeaksPTM. For Paragon, we chose trypsin, Orbi/FT MS (1 ~ 3 ppm) LTQ MS/MS, biological modifications, and the thorough search mode

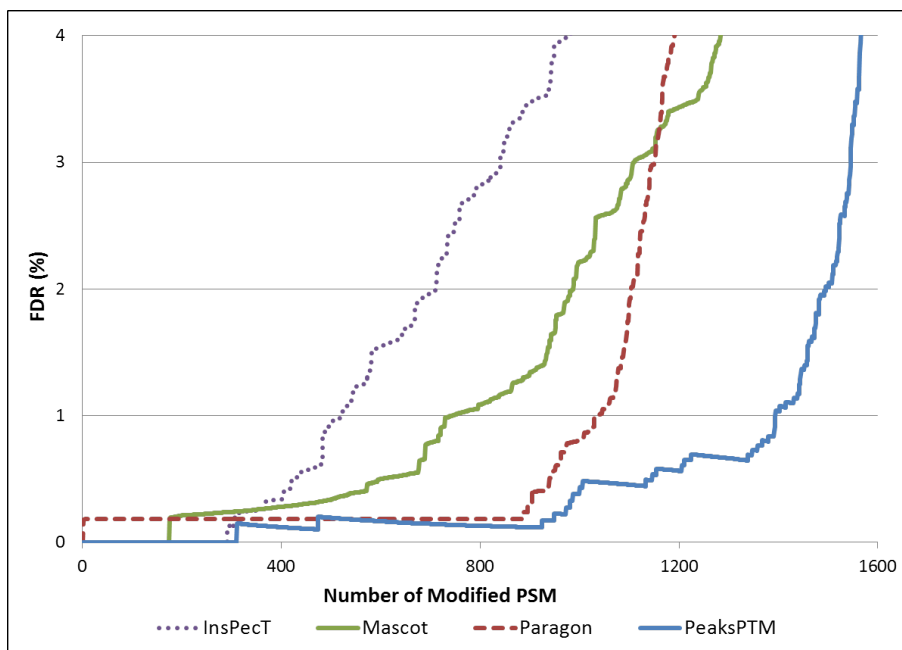


Figure 5.3: The comparison of reported modified PSMs by InsPecT, Mascot, Paragon and PeaksPTM. The curves show the relation between the estimated FDR and the number of modified PSMs.

as the search engine configuration. 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™ i7 CPU with 2.80GHz. InsPecT used 21 CPU hours in total. Using the same computer, PeaksPTM, Paragon and Mascot finished the analysis in approximately an hour, respectively.

With FDR below 1%, PeaksPTM reported 2,410 PSMs, 1,412 of which were modified PSMs; Mascot reported 1,331 PSMs and 729 modified PSMs, Paragon reported 1,972 PSMs and 1,029 modified PSMs, and InsPecT reported 1,133 PSMs and 521 modified PSMs. Figure 5.3 shows the performance comparison of these four software tools on modified PSM identification. 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 5.4. Among the 1,412 modified PSMs, 761 (53.9%) were supported by at

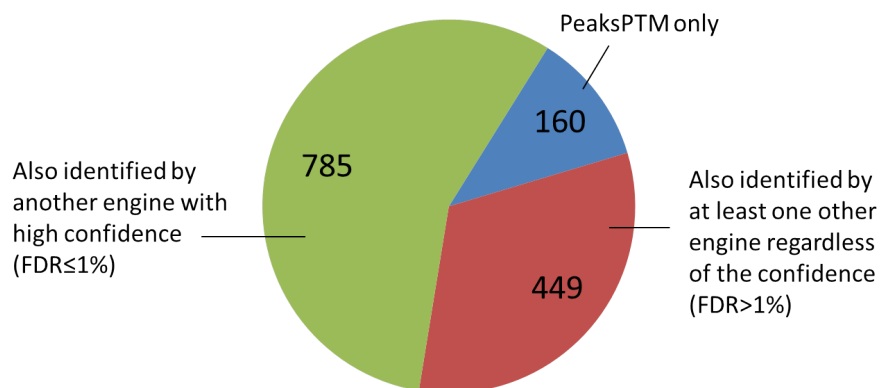


Figure 5.4: A large portion of modified PSMs reported by PeaksPTM with high confidence (FDR ≤ 1%) are also identified by at least one other engine, either with high or low confidence.

least one other search engine with high confidence (with FDR ≤ 1%). 449 (31.8%) additional PSMs were supported by at least one other search engine regardless of the confidence. As it is rare for two engines to falsely identify the same modified PSMs, these consensus identifications are of high confidence.

5.2.4 Comparison with MODⁱ

MODⁱ identifies peptide sequences from a small database containing only at most twenty proteins; therefore, ten top-scoring non-homologous proteins (out of the 1,349 target proteins from the first round search using PEAKS) and their shuffled sequences were combined as the reduced protein database for MODⁱ. All the PTMs provided by the MODⁱ web server were chosen as variable modifications and its default setting for modified mass range (−150 ~ 250 Da) was used. 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. Thus, InsPecT was also added to the comparison with MODⁱ. For a fair comparison InsPecT and PeaksPTM were both used to search the same reduced protein database as MODⁱ.

Figure 5.5 shows the comparison of these three software tools. PeaksPTM still performs best in terms of modified PSM identification. It is noticeable that the FDR

curves can only be used for the purpose of comparing these three tools, but may not accurately reflect the real FDR values of the identifications because of the small size of the target and decoy protein lists.

5.2.5 Consensus Strategy and Analysis

A consensus strategy can be applied to combine the identifications from multiple search engines. A PSM is identified by either more than one search engine with $\text{FDR} \leq 1\%$ or only one search engine with $\text{FDR} \leq 0.8\%$ is considered as a confident identification.

Using this consensus strategy, 3,220 PSMs, including 1,965 modified PSMs, were reported in total by these four search engines. The composition of these 1,965 modified PSMs contributed by four search engines is illustrated in Figure 5.6. Two modified peptides identified by different engines from the same spectrum are regarded as the same if they have the same base form peptides, number of PTMs, and PTM mass shifts. The determination of PTM sites was not considered in this consensus study. The Venn diagram indicates that a large number (871) of modified PSMs were identified by two or more engines confidently and independently. This means that over 36% of all PSMs identified by any single search engine are modified PSMs. The large portion of modified PSMs confirms the belief that the inefficiency in modified peptide identification is one of the major factors for the low identification rate of the MS/MS spectra [146] and the low characterization rate of the modified peptides [39].

5.2.6 Summary of Identified PTMs

Table 5.3 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. There are 906 unique modified peptides identified by PeaksPTM. Oxidation is the most frequent PTM, occurring on 200 peptides. The utilization of the high resolution mass spectrometer enables PeaksPTM to identify PTMs with small Δm , such as deamidation ($\Delta m = 0.98$ Da), but it is still possible that a PTM is mistakenly regarded as another one with the same or very similar Δm .

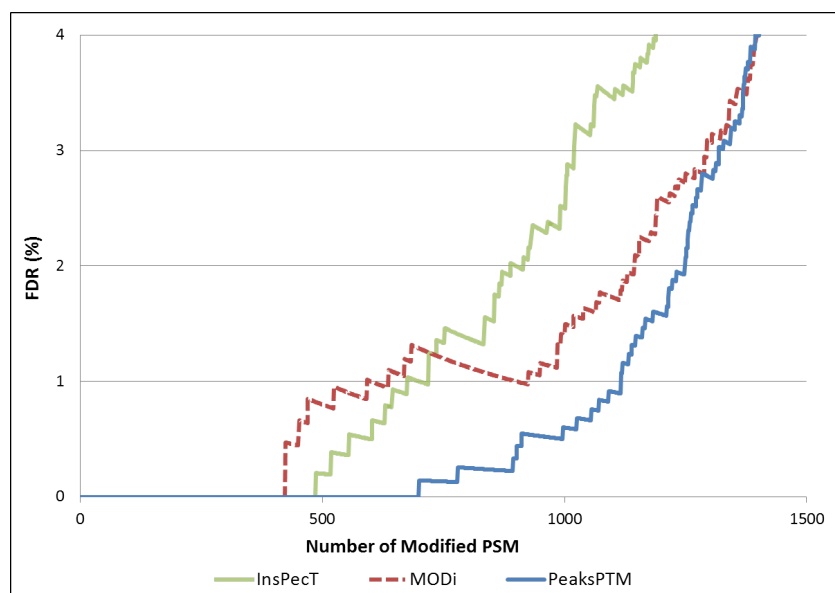


Figure 5.5: The comparison of PeaksPTM, MODⁱ and InsPecT on the reduced database with twenty proteins (ten target and ten decoy proteins). The curves show the relation between the estimated FDR and the number of modified PSMs reported.

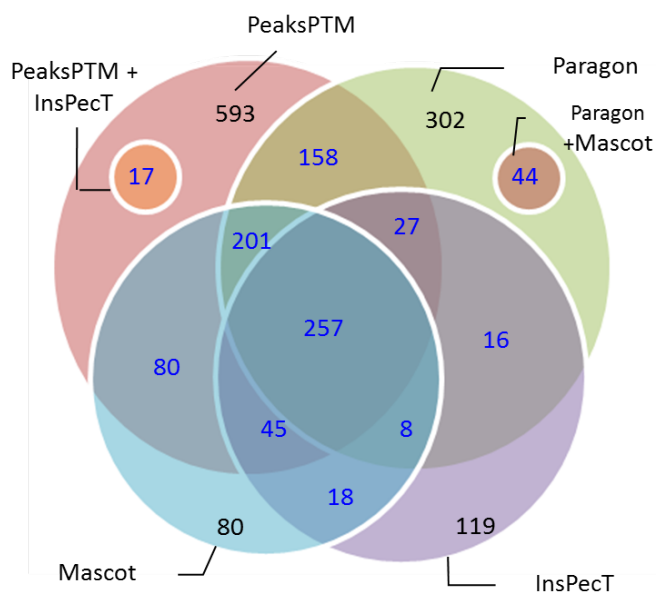


Figure 5.6: The Venn diagram shows the modified PSMs reported by applying the consensus strategy on the results of four search engines.

Table 5.3: The number 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

5.3 Discussion

This chapter proposes our improved database search tool, PeaksPTM, used for modified peptide identification without specifying PTMs. PeaksPTM uses three features, the PEAKS LDF score, the peptide pair, and the PTM rareness, to evaluate modified PSMs. The peptide pair feature is more important according to our statistical analysis: 86.6% of the modified PSMs confidently identified by PeaksPTM have peptide pairs. Compared to using the PEAKS LDF score alone, adding the peptide pair feature and the PTM rareness feature could identify 608 (35.9%) and 156 (9.2%) more PSMs with $FDR \leq 1\%$, respectively, and adding both features improved 717 (42.4%) identified PSMs.

The maximum allowed PTM number was set as 2 to search for the multi-PTM peptides, with consideration of the PTMs summarized in Table 5.3. Only 32 new modified PSMs were identified with high confidence ($FDR \leq 1\%$), while the running time increased up to 3 hours. This experiment demonstrates that (1) the Human-heart data set contains few heavily modified peptides, and (2) the time spent on database searching with multiple PTMs is not negligible, even if only several variable PTMs are considered. Only the single-PTM peptides were compared with the identifications of other search engines in the experimental section.

PeaksPTM is not a blind-search engine that also attempts to find novel PTM types, such as InsPecT. Using all PTM types in the Unimod database is sufficient for most current proteomics research. In our experiment, InsPecT is able to identify only one PTM with mass shift that does not match any existing PTM in the Unimod database. Such identification definitely deserves a careful examination before it is confirmed as a novel PTM. We recommend researchers choose different tools according to their specific applications.

The target-decoy approach widely used today (and also used in PeaksPTM) can only control the false positives at the peptide level, but not at the PTM level, which includes the PTM identity and its location. Consequently, all the FDRs reported in this study measure the correctness of the modified peptide sequence and the Δm of the PTM, but cannot ensure the correctness of the PTM sites reported by those software tools. Accurately locating the PTM site is another nontrivial open problem. It is commonly reckoned that the combination of different types of MS/MS data or the significant improvement of instrument performance will facilitate the progress on this research topic.

Chapter 6

Identification of N-linked Glycopeptides by Tandem Mass Spectrometry

Glycosylation is an enzymatic process that attaches glycans to proteins, lipids, or other organic molecules. It is one of the most frequently observed PTMs and more than 50% eukaryotic proteins are predicted to be glycosylated [6]. Glycosylation of proteins provides either specific structural function induced by conformation changes, or specific recognition sites which are vital to cell-cell interactions [154, 161]. Additionally, increasing evidence suggests that some abnormal glycosylation is strongly correlated with many diseases, such as cancer [80] and congenital disorders [47]. It is commonly believed that these glycan-involved biological processes are closely related to specific glycan structures [32, 112, 153]. Thus, the accurate characterization of glycoproteins, including the amino acid sequences, glycan structures (or composition), and glycosylation sites, is of great interest in the emerging glycoproteomics field [117, 140].

Glycoproteome analysis is more challenging compared with conventional proteome analysis, due to the variety of glycan structures and the complex linkages to proteins. Three types of glycans have been reported: *N*-linked, *O*-linked and *C*-linked. Among the three types, *N*-linked and *O*-linked are the most commonly observed ones. *N*-linked glycans are dominantly found on the Asn residue within a consensus peptide sequence, -Asn-Xxx-Ser/Thr-, where Xxx is any amino acid residue except Pro [51]. Furthermore, *N*-linked glycans share a single core structure, GlcNAc₂Man₃, derived

from the same precursor $\text{GlcNAc}_2\text{Man}_9\text{Glc}_3$ [70]. In contrast, *O*-linked glycans have more varied core structures. This study focuses on the analysis of *N*-linked glycopeptides, including the identification of glycan composition as well as peptide sequences.

Tandem mass spectrometry (MS/MS) is the most powerful tool for the analysis of the glycoproteome because of its high sensitivity and selectivity. In one approach, glycopeptides are deglycosylated partially or totally using a specific glycosidase, such as peptide *N*-endoglycosidase F (PNGase-F), and the resultant peptides and glycan moieties are then analyzed separately by mass spectrometry [58, 59, 147, 167]. Although this method simplifies the data interpretation, it is nontrivial to locate the glycosylation sites for the identified glycans. Several strategies were proposed to characterize intact glycopeptides by MS/MS experiments [103, 117, 162]. In earlier experiments only one type of fragmentation method was used, typically CID, to produce MS/MS spectra. Shortly afterwards, strategies of using a combination of different fragmentation methods for intact glycopeptide analysis emerged. As mentioned in Chapter 2, CID and HCD mainly result in fragment ions through breaking the glycosidic bonds, while ETD and ECD dominantly produce fragment ions by breaking the peptide backbone but leaving the attached glycan intact. Combining two complementary fragmentation techniques in MS/MS analysis enables the identification of peptide sequences, glycan composition, as well as the glycosylation sites [3, 128, 132, 137].

Development of algorithms for automatically interpreting spectra acquired from intact glycopeptides remains in its infancy [5, 126]. GlycoMod [26] is a web-based tool to calculate all possible glycan compositions for a given mass. It does not use the MS/MS data in the analysis. Glyco-Peakfinder [99] and GlycoFragments [92, 93] can calculate the theoretical fragment ions of a given glycan structure, and use them to annotate an MS/MS spectrum. However, these tools cannot identify glycopeptides automatically and require a human expert to deduce the glycan structure. GlycosidIQ [74], GlycoSearchMS [93] and GlycoWorkBench [22] accept a spectrum, search glycan databases, and annotate the spectrum using glycan fragments. A peptide sequence has to be provided to these software tools in advance. This severely limits the capacity for large-scale data analysis. Peptonist [53], which is an extension of Cartoonist [54], and GlypID 2.0 [104] search theoretical glycan structure databases instead of a real database that comprises experimentally validated *N*-linked glycans. Biologically validated rules are used for the generation of the theoretical glycan structures, but not all glycan structures reported in existing glycan databases are covered.

GlycoPep Grader [160] and GlycoPep Detector [170] are web-based tools for assigning the compositions of *N*-linked glycopeptides. However, only one MS/MS spectrum can be processed at a time and users have to input the possible candidate compositions for glycopeptides and glycans. Other software tools, such as STAT [50], Oscar [87], StrOligo [42], GlycoMaster [134], and GLYCH [145], attempt to deduce the glycan structures directly from MS/MS spectra using *de novo* sequencing approaches. These tools typically require spectra with much higher quality for a reliable analysis. This may potentially leave many spectra with medium quality un-interpreted in a high-throughput experiment. A recent review by Dallas *et al.* discussed the current state of glycopeptide assignment software in details and also pointed out the lack of software that could analyze spectral data in batch for the unambiguous characterization of *N*-linked glycopeptides [33].

The protein sequence databases commonly used in proteomics research seldom record the glycan structure information for glycosylated proteins. For example, only 4,375 (21.6%) out of the 20,258 human proteins in the UniProt database contain glycosylation site information. The percentage of glycoproteins is much lower than expected. Thus it is nearly impossible to identify glycopeptides by searching a protein database alone. On the other hand, databases for isolated glycan structures have recently become available, such as CCSD/CarbBank [38, 131], CFG database [124], EUROCarbDB [155], GLYCOSCIENCES.de [95], KEGG [76, 62] and GlycomeDB [125]. Therefore, it is theoretically possible to search a protein sequence database and a glycan structure database simultaneously to characterize the glycopeptides from the spectral data.

In this study, we implement a new software tool, GlycoMaster DB, for the automated and high-throughput characterization of intact *N*-linked glycopeptides from MS/MS data generated by HCD/ETD or HCD-only fragmentation. The software takes MS/MS spectra as input, searches in a given protein sequence database and an integrated glycan structure database simultaneously, and reports the optimal peptide-glycan pair that best matches each spectrum. Performance evaluations on four data sets demonstrate the promising utility of the software.

6.1 Methods

GlycoMaster DB processes MS/MS data from intact glycopeptides. Glycopeptides can be fragmented by either HCD/ETD or HCD-only fragmentation. The HCD/ETD protocol is preferred since the ETD spectra can be used to precisely identify glycopeptide sequences.

A short list of protein sequences needs to be specified by users in a FASTA file. If the glycoproteins are not enriched or enriched at the protein level, a large number of non-glycosylated peptides will be fragmented. Conventional database search tools, such as PEAKS [166], Mascot [118] or Sequest [41], can identify the possible proteins from these non-glycosylated peptides. If the enrichment is performed at the peptide level, the proteins can be identified through separate experiments. The list of proteins provided to GlycoMaster DB can be a mixture of glycosylated and non-glycosylated proteins.

GlycoMaster DB integrates an *N*-linked glycan database extracted from the GlycomeDB database. If required, users can also easily append their own glycans data into this database.

The GlycoMaster DB software analyzes the data in following three steps: (1) filtration of glycopeptide spectra, (2) glycan assignment, and (3) peptide identification. HCD spectra are used in the first two steps for glycan identification. The third step determines the peptide sequences using either ETD data (if available) or the calculated mass values of the peptides bearing the glycan.

6.1.1 Filtration of Glycopeptide Spectra

The input MS/MS data contains a mixture of spectra from both glycosylated and non-glycosylated peptides if the sample is enriched on the protein level or not enriched. GlycoMaster DB first selects out the spectra of glycosylated peptides since this can help to improve the search speed and reduce false positives in later steps.

HCD spectra generated from *N*-linked glycopeptides have two types of characteristics that are not frequently observed in the spectra of non-glycosylated peptides. First, most spectra of *N*-linked glycopeptides have two diagnostic peaks at m/z 204.09 and 366.14, corresponding to oxonium ions formed by a HexNAc and a disaccharide Hex-HexNAc, respectively. Secondly, peaks of a glycopeptide form ion ladders in the

high m/z region. The m/z values of two adjacent singly charged peaks in a ladder differ by the mass of a monosaccharide residue, rather than the mass of an amino acid. Both types of characteristics are used in the algorithm to select the probable glycopeptide spectra. For each spectrum, the diagnostic peaks are first checked. The presence of these two peaks triggers the subsequent examination on the existence of peak ladders. By default, the spectrum is regarded as a glycopeptide spectrum only if it has both the diagnostic peaks and a peaks ladder of length at least four (corresponding to a sequence of three monosaccharide residues) in GlycoMaster DB. Users can also specify the m/z values of diagnostic peaks and the minimum length of monosaccharide ladders. Some glycopeptides only carry one HexNAc modification and our filter will prevent the further analysis on such species. We argue that such single glycosylation can be easily identified by setting HexNAc as a variable PTM in conventional database search software packages.

We design a dynamic programming algorithm to compute the longest sequence of monosaccharide residues that matches a series of high-intensity peaks in the spectrum. In the algorithm, all the mass values are converted to the equivalent nominal mass by multiplying a factor 0.9995 and then rounding to the nearest integers [14, 67]. After the conversion, we select the highest 50 peaks of the spectrum to calculate the longest sequence.

In a preprocessed spectrum, a sequence of monosaccharide residues is represented by a series of peaks at m/z values m_1, \dots, m_{k+1} , where $(m_{i+1} - m_i), i \in [1, \dots, k]$, is equal to the mass of a monosaccharide residue. The length of such a sequence is k . The *longest sequence of monosaccharide residues (LSMR)* problem is to find the maximum value of k in a given spectrum. Three most frequently observed monosaccharide residues, Hex, HexNAc, and Fuc, are considered in this algorithm as the residue set. Let $L[m]$ be the length of the longest sequence that ends at mass m , and $L[m] = -1$ if there is no peak at mass m . If a peak is present at mass m , the algorithm needs to check the existence of a shorter sequence ends at mass m_0 such that $(m - m_0)$ is equal to a monosaccharide residue mass. Such checking needs to be carried out for each of the three given residues. Therefore, $L[m] = \max_{i \in [1,3]} L[m - m(r_i)] + 1$, where $m(r_i)$ is the mass of the i -th monosaccharide residue r_i . To summarize, the algorithm MaxSeqLen shown in Algorithm 2 can compute $L[m]$ for every mass value m . The running time is linear to the precursor mass of the spectrum.

Algorithm 2 The MaxTagLength algorithm for solving the longest sequence of monosaccharide residues (LSMR) problem.

Require: An MS/MS spectrum S .

```
1: function MAXTAGLENGTH( $S$ )
2:   Let  $T[0] \leftarrow -1$ 
3:   Let  $M \leftarrow$  the precursor mass of the spectrum  $S$ 
4:   for  $m \leftarrow 1$  to  $M$  do
5:     if there is no peak at  $m$  then
6:        $T[m] = -1$ 
7:     else
8:        $T[m] = \max_{i \in [1,3]} T[m - m(r_i)] + 1$ 
9:   return  $\max_{m \in [1,M]} T[m]$ 
```

6.1.2 Glycan Assignment

If a spectrum is regarded as a possible glycopeptide spectrum, the N -linked glycan database is searched for its best matching glycan. Glycans that have smaller mass than the precursor mass of the spectrum are matched to the spectrum. Each *glycan-spectrum match* (GSM) is evaluated and the glycan with the highest score is reported. The GSM scoring scheme is designed similarly to the ones commonly used for peptide identification: (1) the theoretical m/z values of the possible fragment ions are calculated, (2) for each fragment ion, a reward or a penalty is added to the score depending on whether its m/z value matches a peak in the spectrum. These two components are described in the following two subsections, respectively.

Glycan Structure Fragmentation

As shown in Figure 2.4, HCD favors the fragmentation of glycosidic bonds rather than the peptide bonds and produces B -, Y -, C -, and Z -ions [37]. In theory, a breakage can also occur across the ring of a monosaccharide to produce A - and X -ions. However, in practice, Y -ions are the most commonly observed ions in HCD spectral data. Furthermore, peaks representing oxonium ions and B -ions can be observed in the low m/z region, and in most cases, only those product ions with at most three monosaccharide residues generate significant peaks. Therefore, B - and oxonium ions with at most three monosaccharide residues, as well as Y -ions, are considered in our

scoring scheme. GlycoMaster DB takes a condensed GlycoCT file in the GlycomeDB database as input, parses it into a tree structure, and enumerates all the expected *B*-, *Y*- and oxonium ions as discussed above.

The theoretical m/z values of the ions are calculated during the ion enumeration. For example, for singly charged ions, the m/z value of a *B*- or oxonium ion is equal to the total mass of the monosaccharide residues plus an additional proton, and the m/z value of a *Y*-ion is equal to the singly charged precursor m/z value subtracting the mass of the removed monosaccharide residues. The list of theoretical m/z values and their corresponding fragment ion types are provided to our scoring scheme for GSM evaluation.

Glycan-Spectrum Matching Score

In contrast with the development of PSM score in proteomics, the main challenge for developing the scoring scheme for GSM is the lack of large-scale training data. The proper values of the reward and penalty for a fragment ion matching and mismatching may depend on many factors such as the fragment ion type, the intensity of the matching peak, and the mass error. In proteomics, these values are usually statistically learned from a large number of training spectra annotated with known results. Unfortunately, in the glycoproteomics field, such a large-scale training data set is not yet available. Therefore, an empirical scoring function is used.

The scoring scheme in GlycoMaster DB calculates raw scores of GSMs first. Given a glycan structure and a spectrum, the theoretical m/z values of the glycan fragment ions are searched in the spectrum. The score S for a fragment ion matched by a peak with relative intensity I is calculated using the following equation:

$$S = \begin{cases} \lg(100 \times I), & \text{If a peak with relative intensity } I \geq 0.5\% \text{ is matched} \\ \lg 0.5, & \text{Otherwise} \end{cases} \quad (6.1)$$

The GSM raw score is the sum of all fragment ion scores. The glycan structure with the highest GSM raw score is reported as the best match for the given spectrum.

The GSM raw score serves the purpose of selecting the best matching glycan structure since a correct structure often produces more high-intensity matches and generates a higher score than false structures. However, an incorrect GSM of a spectrum with numerous peaks can easily get a higher raw score than a correct GSM

of a spectrum with few peaks. Therefore, to compare the GSMs of different spectra, the raw score is further normalized to a $-10 \lg P$ score, where P denotes the p -value.

A $-10 \lg P$ score represents the confidence of a GSM. Given a spectrum, the raw scores of all glycans in the database are used to fit a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where μ and σ are the mean and the standard deviation of the GSM raw scores, respectively. Each raw score x is used to compute a p -value P that denotes the probability in which a random variable under $\mathcal{N}(\mu, \sigma^2)$ exceeds x . The final GSM score is $-10 \lg P$ and displayed in the result reported by GlycoMaster DB. The identification results are sorted according to GSM scores. In our study, the reported glycan of which the GSM score is no less than 15 (corresponding to a p -value of 3.2%) is regarded as the plausible identification of a spectrum.

6.1.3 Glycopeptide Identification

GlycoMaster DB accepts two types of MS/MS data as input: HCD/ETD spectrum-pairs and HCD-only spectra. Therefore, two different approaches for glycopeptide identification were implemented separately.

Peptides cannot be identified from HCD spectra since few fragment ions from the backbones are generated. In contrast, the ETD method dominantly produces fragment ions by breaking a peptide backbone but leaving the attached glycan intact. Thus, peptides are identified from ETD spectra in GlycoMaster DB when HCD/ETD spectrum-pairs are available. The peptides containing the N -linked glycopeptide motifs are generated first using the user-specified enzyme. For an ETD spectrum, a peptide containing an N -linked glycopeptide motif and with mass smaller than the spectrum's precursor mass is considered as a candidate, and the mass difference is regarded as the mass of the glycan. The peptide backbone fragment ions of each glycopeptide candidate are then matched to the ETD spectrum to calculate a PSM raw score. The Eq. 6.1 is used in the PSM raw score calculation with consideration of c -, c -H, z -, z' - and z'' -ions. This raw score is then converted into a $-10 \lg P$ score as the final PSM score, using the same procedure of the GSM score calculation. Then, for each HCD/ETD spectrum-pair, the glycans obtained from the HCD spectrum and the peptides from the ETD spectrum are combined together to build glycopeptides. A glycopeptide is regarded as a probable identification to a spectrum if (1) the precursor mass error is within the allowed mass error tolerance and (2) either the GSM score or

the PSM score is greater than or equal to 15. If multiple glycopeptides satisfy these two criteria, the one with the highest GSM score is kept in the main report, and the others are stored in a secondary table that can be further examined by users. If no peptide sequence is found for a spectrum-pair, the glycan with the top GSM score from the HCD spectrum and a calculated peptide mass are reported.

If only HCD spectra are available, the peptide sequences are identified from two sources of information: the calculated mass of the peptide and the existence of an *N*-linked glycopeptide motif. The peptides containing such motifs are generated first using the user-specified enzyme and stored in a sorted list in ascending order of mass. GlycoMaster DB identifies glycans from an HCD spectrum and the glycan with top GSM score (at least 15, which corresponds to a *p*-value of 3.2%) is kept for later peptide determination. The difference between the precursor mass of the spectrum and the mass of the top-scored glycan is the peptide mass. A binary search is then applied to find peptides with this mass from the peptide list. The resultant glycopeptides matching the spectrum precursor mass within the mass error tolerance are reported as a list of possible peptides for the spectrum. If no peptide is found, only the calculated peptide mass is reported.

6.2 Results

Four previously published data sets (Ribonuclease B, Human Immunoglobulin G, Lectin-Enriched Human Urinary Proteome, and Human Urinary Proteome) by Singh *et al.* [137] and Marimuthu *et al.* [103] were used to evaluate the performance of GlycoMaster DB. The first two data sets were obtained with HCD/ETD fragmentation and thus glycopeptides could be characterized both on glycan composition and peptide sequences. For the human urinary proteome data sets obtained with HCD fragmentation, GlycoMaster DB identified the glycan composition, while the peptide sequences were reported only according to the calculated masses. Clearly, several peptides may share the same mass value, resulting in peptide identification ambiguities if HCD-only data is used. To study the severity of this ambiguity, statistical analysis by computational simulation was conducted and its results are illustrated at the end of this section.

Experimental procedures for the sample preparation, glycoprotein enrichment and LC-MS/MS analysis were described in details in Singh *et al.* [137] and Marimuthu *et*

al [103]. Here, we briefly introduce the four data sets as follows:

Ribonuclease B (RNase-B) Data Set: This data set was from the study of HCD product ion-triggered ETD (HCD PI ETD) analysis for characterization of glycoproteins proposed by Singh *et al* [137]. Ribonuclease B (RNase B) from bovine pancreas was digested using Lys-C. The digested peptides were separated using a zwitterionic hydrophilic interaction liquid chromatography nano-column, and then analyzed using LTQ-Orbitrap Velos (Thermo Fisher Scientific, Bremen, Germany). The mass spectrometer performed a full survey scan with Orbitrap and subsequent HCD MS/MS scans of the 40 most abundant ions. If peaks at m/z 204.09 (HexNAc oxonium ions) or 366.14 (Hex-HexNAc oxonium ions) ($\pm m/z$ 0.05) were within the top 20 most abundant peaks, a supplemental activation ETD MS/MS scan of the precursor ion in the linear ion trap was triggered. This data set contained 3,111 MS spectra and 774 MS/MS spectra (632 HCD and 142 ETD spectra).

Human Immunoglobulin G (Human-IgG) Data Set: This data set was from HCD PI ETD analysis for characterization of glycopeptides in human IgG proteins [137]. Human IgG is an antibody isotype and its fragment crystallizable region bears a highly conserved *N*-linked glycosylation site. Four subclasses of human IgG, IgG1, IgG2, IgG3, and IgG4, were present in this analysis. These proteins were digested by trypsin and analyzed with the same HCD PI ETD strategy used for the acquirement of the RNase-B data set. This data set comprised 952 MS spectra and 5,710 MS/MS spectra (5,436 HCD and 274 ETD spectra).

Lectin-Enriched Human Urinary Proteome (Enriched-HUP) Data Set: This data set was from a comprehensive analysis of human urine proteome by Marimuthu *et al.* [103] and contained 24 raw data files. The sample was incubated with a mixture of three agarose conjugated lectins – concanavalin A, wheat germ agglutinin and jacalin (Amersham BioSciences) – for glycoprotein enrichment. The concentrated protein was then resolved by SDS-PAGE and visualized using colloidal Coomassie staining. Twenty-four bands were excised and subjected to in-gel trypsin digestion procedure and then analyzed using LTQ-Orbitrap Velos (Thermo Fisher Scientific, Bremen, Germany) interfaced with an Agilent’s 1200 Series nanoflow LC system. The mass spectrometry analysis was carried out in a data dependent mode with survey scans acquired using Orbitrap mass analyzer, and 20 most abundant precursor ions from a survey scan were selected for HCD MS/MS scans. This data set contained 22,886 MS spectra and 199,890 MS/MS spectra in total.

Human Urinary Proteome (HUP) Data Set: This data set was also from the comprehensive analysis of human urinary proteome and included 30 raw data files. The sample was separated by SDS-PAGE without lectin-enrichment of glycoproteins. Thirty gel bands were excised and subjected to in-gel trypsin digestion. The sample analysis was carried out as described in the Enriched-HUP data set. This data set included 35,788 MS spectra and 170,215 MS/MS spectra in total.

All four data sets were analyzed using PEAKS to identify the lists of proteins with $FDR \leq 1\%$. The resultant proteins were exported as FASTA files for GlycoMaster DB analyses. The RNase-B data set was searched against the UniProt bovine database (5,973 entries), and the Human-IgG data set was searched against the UniProt human database (20,258 entries). Oxidation of Met was set as a variable PTM and carbamidomethylation of Cys as a fixed PTM. The maximum allowed number of missed-cleavages was set to two. The precursor and fragment error tolerances were 10 ppm and 0.1 Da, respectively. The two human urinary proteome data sets were searched against UniProt human database (20,258 entries). Oxidation of Met, deamidation at Asn and Gln, and protein N-terminal acetylation were selected as variable PTMs and carbamidomethylation of Cys as a fixed PTM. One missed-cleavage was allowed for tryptic peptides. The precursor and fragment error tolerances were 20 ppm and 0.1 Da, respectively.

In subsequent GlycoMaster DB analyses, these four data sets were searched against our integrated *N*-linked glycan database containing 2,925 unique *N*-linked glycans. These glycans were extracted from the GlycomeDB database and two glycans were regarded as the same if they had the same tree structure but different linkages between the monosaccharide residues. For each data set, the PTMs, the mass error tolerances of precursor and fragment ions, and the maximum number of missed-cleavage were set the same to the ones used in PEAKS analysis. Sodium and potassium adducts were also considered in the search.

6.2.1 RNase-B Data Set

This data set was obtained using the HCD PI ETD strategy. HCD spectra were preprocessed by the Data Refine module in PEAKS and thereafter used to identify the short list of proteins. Among the 774 MS/MS spectra, 31 were identified as non-glycosylated peptides with high confidence ($-10 \lg P \geq 34.4$ and $FDR \leq 1\%$)


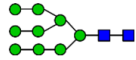
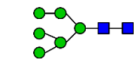
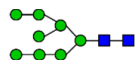
No.	Scan	Precursor m/z	Precursor Charge	RT	Glycan ID	Glycan	GSM Score	Peptide	PSM Score	Error (ppm)	Protein ID
1	2340	807.672	3	28.09	12572		50.72	SRN(+1702.59)LTK	104.84	-0.38	P61823
2	2437	861.6897	3	28.95	12573		50.48	SRN(+1864.64)LTK	91.33	-0.57	P61823
3	2467	699.6374	3	29.21	13764		50.18	SRN(+1378.48)LTK	33.38	-0.96	P61823
4	2447	807.6714	3	29.04	12572		46.89	SRN(+1702.59)LTK	63.89	0.3	P61823

Figure 6.1: An example of the GlycoMaster DB result page generated in an HCD/ETD spectral data analysis. The results are listed in a HTML table in descending order of glycan scores. Each row represents an identification of an HCD/ETD spectrum pair. The first column includes a hyperlink that redirects to the top-ten interpretations of the same HCD/ETD spectrum-pair. The second to the fifth column list the spectrum information. The sixth and the seventh column list the glycan information obtained from the GlycomeDB database, and the hyperlinks redirect to the GlycomeDB website. The eighth column gives scores of GSMs and each hyperlink redirects to the annotated HCD spectrum and its mass error chart. The ninth and tenth column list the peptide sequences and the PSM scores obtained from ETD spectra. The hyperlink at the PSM score column links to the annotated ETD spectrum and the mass error chart. The mass error between the theoretical and experimental mass values of an identified glycopeptide is provided in the eleventh column. The last column gives the accession numbers of corresponding proteins.

Table 6.1: The grouped results identified by GlycoMaster DB from the RNase-B data set. The spectra with the same precursor m/z and charge are grouped in a row if they have the same identification. The GSM score, PSM score and mass error in a row are from the HCD/ETD spectrum-pair with the highest GSM score.

Precursor m/z	Precursor Charge	RT Range	Glycan Composition	GSM Score	Glycan Mass	PSM Score	Error (ppm)
886.8996	2	27.06-28.21	HexNAc ₂ Hex ₄	33.2	1054.37	55.2	-0.89
967.9255	2	20.13-29.67	HexNAc ₂ Hex ₅	41.34	1216.43	73.44	-0.44
645.6194	3	19.79-28.54	HexNAc ₂ Hex ₅	39.72	1216.43	45.04	-0.09
699.63715	3	24.74-29.21	HexNAc ₂ Hex ₆	46.06	1378.49	52.65	-0.61
1048.9491	2	27.33-29.35	HexNAc ₂ Hex ₆	36.62	1378.49	52.69	1.98
753.65375	3	27.56-29.59	HexNAc ₂ Hex ₇	42.83	1540.54	40.01	0.65
1129.9783	2	29.27	HexNAc ₂ Hex ₇	36.59	1540.54	82.64	-0.76
767.33167	3	28.31	HexNAc ₃ Hex ₆	32.18	1581.57	34.22	-2.31
807.672	3	26.86-30.52	HexNAc ₂ Hex ₈	50.72	1702.59	72	-0.38
861.6832	3	28.02-28.95	HexNAc ₂ Hex ₉	50.48	1864.64	73.19	-0.57

and nine proteins were reported. HCD/ETD spectrum-pairs were then extracted for glycopeptide analysis using GlycoMaster DB.

142 HCD/ETD spectrum-pairs were collected and 31 of these pairs were identified by GlycoMaster DB. Figure 6.1 illustrates the result page of GlycoMaster DB on the RNase-B data set (only the first four entries of the result are shown). The identified glycans and peptide sequences are listed in a HTML table in descending order of the glycan score. Users can easily check the annotated HCD or ETD spectrum and the glycan information in the GlycomeDB database through the hyperlinks in the result page. The 31 HCD/ETD spectrum-pairs identified by GlycoMaster DB have only 10 unique precursor m/z and charge combinations. Thus their results are grouped and listed in Table 6.1. All these identifications share the single peptide sequence SRNLTK.

Figure 6.2 illustrates an example of a glycopeptide identified from an HCD/ETD spectrum-pair by GlycoMaster DB. Both the HCD spectrum and the triggered ETD spectrum have a same precursor m/z value and similar retention time. Clearly, in the HCD spectrum (Figure 6.2(a)), the peak ladder started from m/z 921.5 is definitely

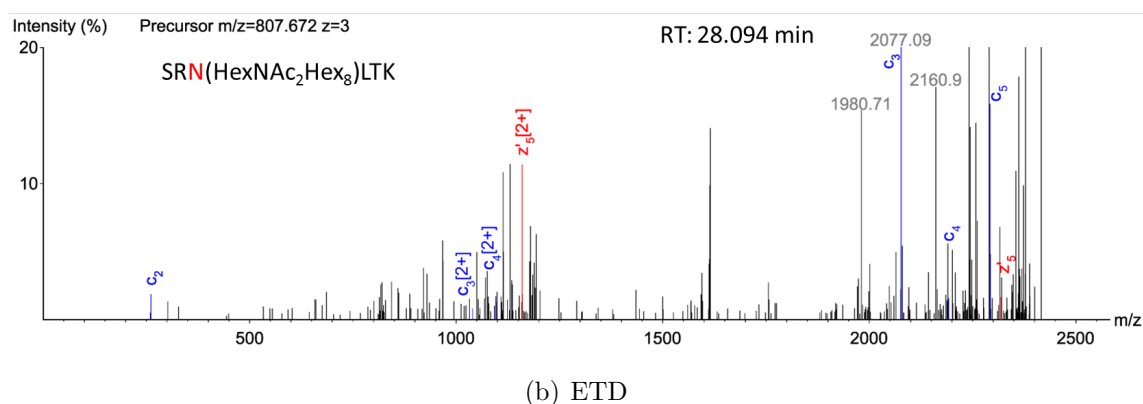
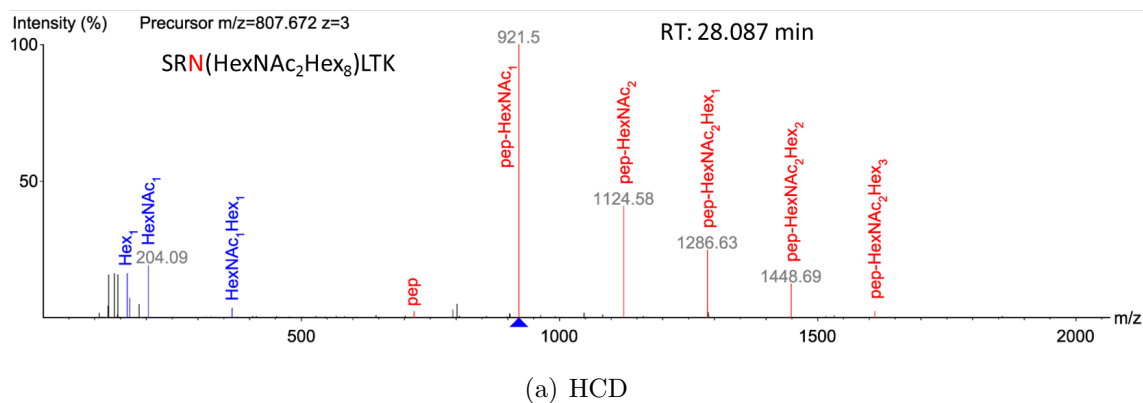


Figure 6.2: An example of a glycopeptide identified from the RNase-B data set generated by the HCD PI ETD strategy. (a) The annotated HCD spectrum of precursor ions with m/z 807.672. GlycoMaster DB identified the best matched glycan with the composition HexNAc2Hex8. SRNLTK is the only potential glycopeptide having the similar mass to the calculated mass 699.404. (b) The annotated ETD spectrum triggered by product ions in the HCD spectrum shown in (a). It provides positive support for the identification of the peptide SRNLTK and the glycosylation site.

from the Y -ions of the glycopeptide. A glycan with the composition HexNAc₂Hex₈ was reported by GlycoMaster DB as the best matching glycan with the highest GSM score. The calculated mass of the peptide has only one peptide SRNLTK matched within the given mass error tolerance in the nine proteins identified by PEAKS. For the ETD spectrum as shown in Figure 6.2(b), GlycoMaster DB separately identified the same peptide with the PSM score 72. Therefore, GlycoMaster DB reported the glycopeptide SRN(HexNAc₂Hex₈)LTK as the identification of this HCD/ETD spectrum-pair. The theoretical triply charged precursor m/z is 807.6717 and it differs from the experimental precursor m/z with only 0.38 ppm.

As an optional step to further validate the peptide sequence, the PEAKS database search software can be used to analyze the ETD spectra. A glycan of which the mass has been determined by GlycoMaster DB is provided to PEAKS as a user-defined variable PTM. PEAKS can check all the *in silico* digested peptides, rather than only the peptides with N -linked glycopeptide motifs. Therefore, if the best matching peptide for the ETD spectrum has an N -linked glycopeptide motif and its PSM score is high, it is regarded as the identification of this ETD spectrum with high confidence. We set all the glycans reported by GlycoMaster DB as variable PTMs for PEAKS database search. Among 142 ETD mass spectra, PEAKS identified 31 spectra with $-10 \lg P \geq 15$ and all of them were identified as the same peptide SRNLTK.

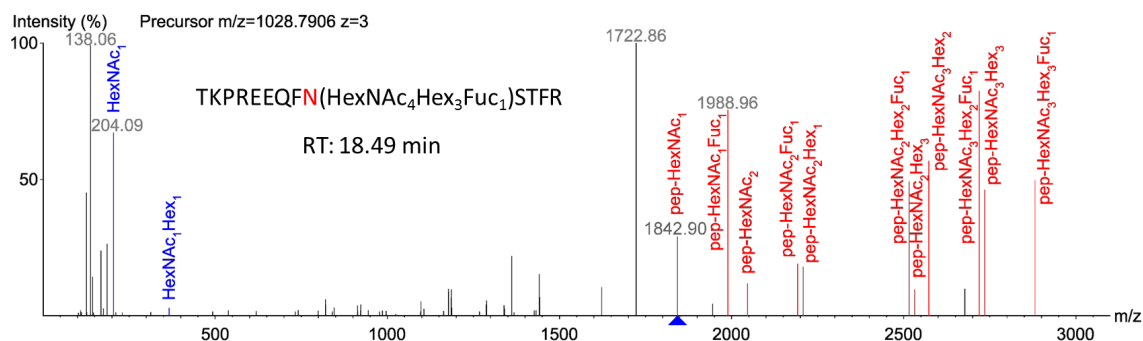
6.2.2 Human-IgG Data Set

Similarly to the analysis of the RNase-B data set, HCD spectra in the Human-IgG data set were used to identify a short list of proteins. Among the 5,710 MS/MS spectra, 306 were identified as non-glycosylated peptides and 36 proteins were reported. HCD/ETD spectrum-pairs were then extracted for glycopeptide analysis using GlycoMaster DB. Out of the 274 HCD/ETD spectrum-pairs, 10 spectrum-pairs were reported with either the GSM score or the PSM score higher than 15. The reported glycopeptides are listed in Table 6.2.

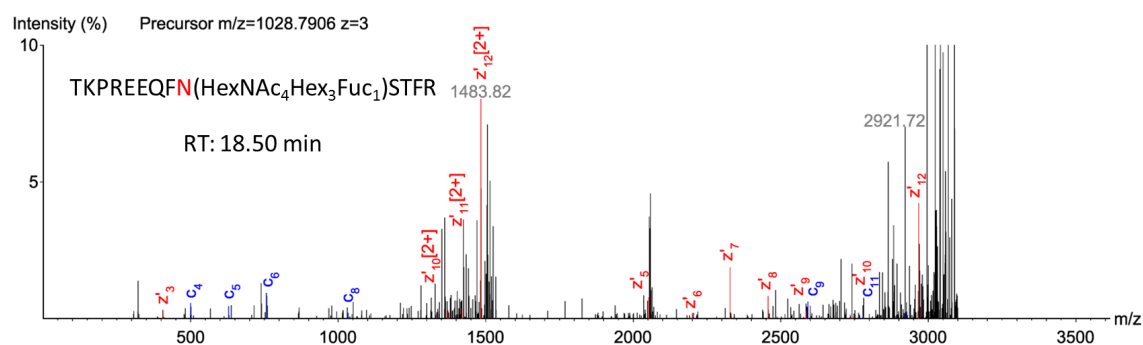
Figure 6.3 illustrates a glycopeptide identified from an HCD/ETD spectrum-pair by GlycoMaster DB. Figure 6.3(a) shows the HCD spectrum recorded at RT 18.49 min. The glycan reported by GlycoMaster DB has the composition HexNAc₄Hex₃Fuc₁, which forms a clear Y -ion ladder in the high m/z region. From the calculated mass of the peptide, TKPREEQFNSTFR is selected as the possible peptide sequence. This

Table 6.2: Glycopeptide identified by GlycoMaster DB from the Human-IgG data set.

Precursor m/z	Precursor charge	RT	Glycan	Glycan Score	Glycopeptide	Peptide Score	Error (ppm)	IgG Subclass
1301.532	2	12.55	HexNAc ₄ Hex ₃ Fuc ₁	27.4	EEQFN(+1444.54)STFR	14.26	-0.28	IgG 2/3
1301.532	2	11.89	HexNAc ₄ Hex ₃ Fuc ₁	19.09	EEQFN(+1444.54)STFR	11.91	-0.75	IgG 2/3
1317.527	2	12.52	HexNAc ₄ Hex ₃ Fuc ₁	25.43	EEQYN(+1444.54)STYR	14.99	-0.56	IgG 1
1028.791	3	18.49	HexNAc ₄ Hex ₃ Fuc ₁	19.39	TKPREEQFN(+1444.54)STFR	42.96	-1.42	IgG 2/3
1082.802	3	18.69	HexNAc ₄ Hex ₄ Fuc ₁	12.92	TKPREEQFN(+1606.59)STFR	53.68	3.83	IgG 2/3
1082.809	3	18.17	HexNAc ₄ Hex ₄ Fuc ₁	10.77	TKPREEQFN(+1606.59)STFR	26.68	-2.25	IgG 2/3
1093.473	3	18.74	HexNAc ₄ Hex ₅ Fuc ₁	10.64	TKPREEQFN(+1622.59)STYR	46.2	-2.46	IgG 1
1039.454	3	19.56	HexNAc ₄ Hex ₃ Fuc ₁	14.42	TKPREEQYN(+1444.54)STYR	52.57	-1.41	IgG 1
1093.473	3	19.32	HexNAc ₄ Hex ₄ Fuc ₁	12.25	TKPREEQYN(+1606.59)STYR	62.64	-2.57	IgG 1
1093.472	3	19.90	HexNAc ₄ Hex ₄ Fuc ₁	10.51	TKPREEQYN(+1606.59)STYR	45.05	-2.01	IgG 1



(a) HCD



(b) ETD

Figure 6.3: An example of a glycopeptide identified from the Human-IgG data set generated by the HCD PI ETD strategy. (a) The annotated HCD spectrum of precursor ions with m/z 1028.7906. The best matched glycan reported by GlycoMaster DB has the composition HexNAc₄Hex₃Fuc₁. (b) The annotated ETD spectrum triggered by product ions in the HCD spectrum shown in (a). It provides positive support for the identification of peptide TKPREEQFNSTFR and the glycosylation site.

peptide is also identified separately by GlycoMaster DB from the ETD spectrum shown in Figure 6.3(b) with the highest PSM score. The difference between the theoretical and the experimental m/z of this glycopeptide is 1.42 ppm.

PEAKS database search on the ETD mass spectra was further carried on to validate the peptide identification reported by GlycoMaster DB. We set all the glycans reported by GlycoMaster DB as variable PTMs for PEAKS database search. Among 274 ETD mass spectra, PEAKS reported 20 PSMs with $FDR \leq 1\%$ and all the peptides identified by GlycoMaster DB with GSM scores of higher than 15 were included. PEAKS also identified nine non-glycosylated peptides, which matched to both HCD and ETD mass spectra with high PSM scores. Manual checking revealed that their HCD spectra had low peaks at m/z 204.09, which falsely triggered the generation of the corresponding ETD spectra. However, these spectrum-pairs for the non-glycosylated peptides did not result in false positives in GlycoMaster DB's results.

6.2.3 Enriched-HUP Data Set

This data set contains a large amount of spectra with duplicated precursor m/z and similar retention time, indicating that many peptides were selected and fragmented multiple times. Therefore, two MS/MS scans in this data set were merged with the PEAKS software if their precursor m/z difference was within 20 ppm and the retention time difference is within 0.2 minutes. The 24 spectral data files were then searched separately in UniProt human protein database for the short lists of proteins through identifying the non-glycosylated peptides and 503 proteins were reported. A spectrum was filtered out if it was identified either by PEAKS DB with $FDR \leq 1\%$ or by PEAKS *de novo* sequencing with $ALC \geq 50\%$ since it was believed to be from a non-glycosylated peptide. The remaining MS/MS spectra were then analyzed by GlycoMaster DB. The results from the 24 spectral data files are listed in Table 6.3.

In total, 14,840 spectra were not interpreted by either the database searching or the *de novo* sequencing modules in PEAKS. These spectra were analyzed by GlycoMaster DB. 2,283 spectra passed the first filtration according to the existence of diagnostic peaks at m/z 204.09 or 366.14, and 451 spectra had sequences of at least three monosaccharide residues. These 451 spectra were searched against the *N*-linked glycan database. 240 spectra had matched glycans with high confidence ($-10 \lg P \geq 15$),

Table 6.3: This table lists the analysis result of each spectral data in enriched-HUP data set. The spectra data named “gpe12” is not listed since it has no glycan reported by GlycoMaster DB. The first column lists the names of the spectral files. The second column denotes the numbers of MS/MS spectra in each file after preprocessing. The third and fourth column lists the numbers of proteins reported by PEAKS DB and the numbers of un-interpreted spectra. The subsequent two columns give the number of spectra that passed the two filters, respectively. The number of identified glycans ($-10 \lg P \geq 15$) and peptides are listed in the last columns. The last row shows the total number of each column.

Data Name	MS/MS Number	Protein Number	Un-Interpreted MS/MS Number	Pass Filter-1	Pass Filter-2	Glycan Number	Peptide Number
gpe01	1,635	53	604	109	12	10	10
gpe02	1,977	63	890	106	15	7	7
gpe03	1,821	65	798	99	17	6	6
gpe04	1,854	53	844	123	37	7	4
gpe05	1,920	59	832	107	31	20	18
gpe06	1,683	68	626	128	36	17	12
gpe07	1,866	100	835	108	14	6	6
gpe08	1,712	83	683	116	14	9	9
gpe09	1,777	77	788	112	10	8	8
gpe10	1,798	87	677	95	15	6	5
gpe11	1,999	86	906	210	12	4	4
gpe13	927	71	250	65	30	21	21
gpe14	929	65	301	84	38	17	13
gpe15	1,101	68	466	101	38	18	15
gpe16	1,290	91	559	68	21	14	14
gpe17	1,227	94	554	81	13	12	12
gpe18	1,282	88	597	93	26	16	16
gpe19	1,333	84	620	104	18	12	10
gpe20	1,234	81	619	106	23	6	5
gpe21	955	44	491	61	16	13	13
gpe22	977	79	482	68	8	6	4
gpe23	770	59	346	31	1	1	1
gpe24	1,117	65	711	90	6	4	3
Total	33,954	503 ^a	14,840	2,283	451	240	216

^aThis is the total number of unique proteins, rather than the sum of protein numbers reported in each spectral data.

and 216 of them had matched peptides by masses. Possible reasons for not reporting peptide sequences for the 25 remaining spectra include (1) the peptides are not in the 503 proteins identified by PEAKS and (2) the peptides may be the result of non-specific trypsin digestion, have more missed-cleavages, or have variable PTMs other than those considered. 95 proteins were reported as glycoproteins with at least one glycopeptide identified by GlycoMaster DB in each protein.

Figure 6.4 shows three example glycopeptides identified by GlycoMaster DB. Figure 6.4(a) shows an MS/MS spectrum recorded at RT 30.03 min. The precursor m/z 1328.0316 corresponds to a doubly charged glycopeptide N(HexNAc₂Hex₉)WTITR ($m/z_{calc} = 1328.0299$ and $\Delta m/z = -1.29$ ppm). The peak at m/z 993.5 corresponds to the singly charged [peptide+HexNAc] ion, which is the Y_1 fragment according to the Domon and Costello nomenclature [37]. Figure 6.4(b) and Figure 6.4(c) show other two HCD spectra identified as having the same peptide sequence but slightly different glycans. The differences between their precursor masses are the masses of monosaccharide residues. As the retention time is mainly determined by the hydrophobicity of amino acids instead of the glycans attached on the glycopeptides, the precursor ions of these spectra have similar retention time.

Figure 6.5 illustrates another example of two similar glycans on the same peptide. The retention time of these two spectra differs by 2.7 min. The identified glycans are very similar and SLHVPGLNK is the only glycopeptide that has the calculated peptide mass. Consequently, the two spectra are very similar to each other, except that Figure 6.5(b) contains two intense peaks at 292.10 and 274.09, which are missing from Figure 6.5(a). These peaks demonstrate the existence of sialic acid residues. It is commonly believed that sialic acids can influence the retention time of glycopeptides. This is consistent with the 2.7 min retention time difference between the two spectra.

6.2.4 HUP Data Set

This data set was obtained from the same human urine sample as the Enriched-HUP data set. The only difference was the glycoproteins were not enriched. GlycoMaster DB was used to process this data set since many spectra contained high-intensity diagnostic peaks of *N*-linked glycopeptides.

Similarly to the analysis of Enriched-HUP data set, the 30 spectral data files were searched separately in UniProt human protein database for the short lists of proteins.

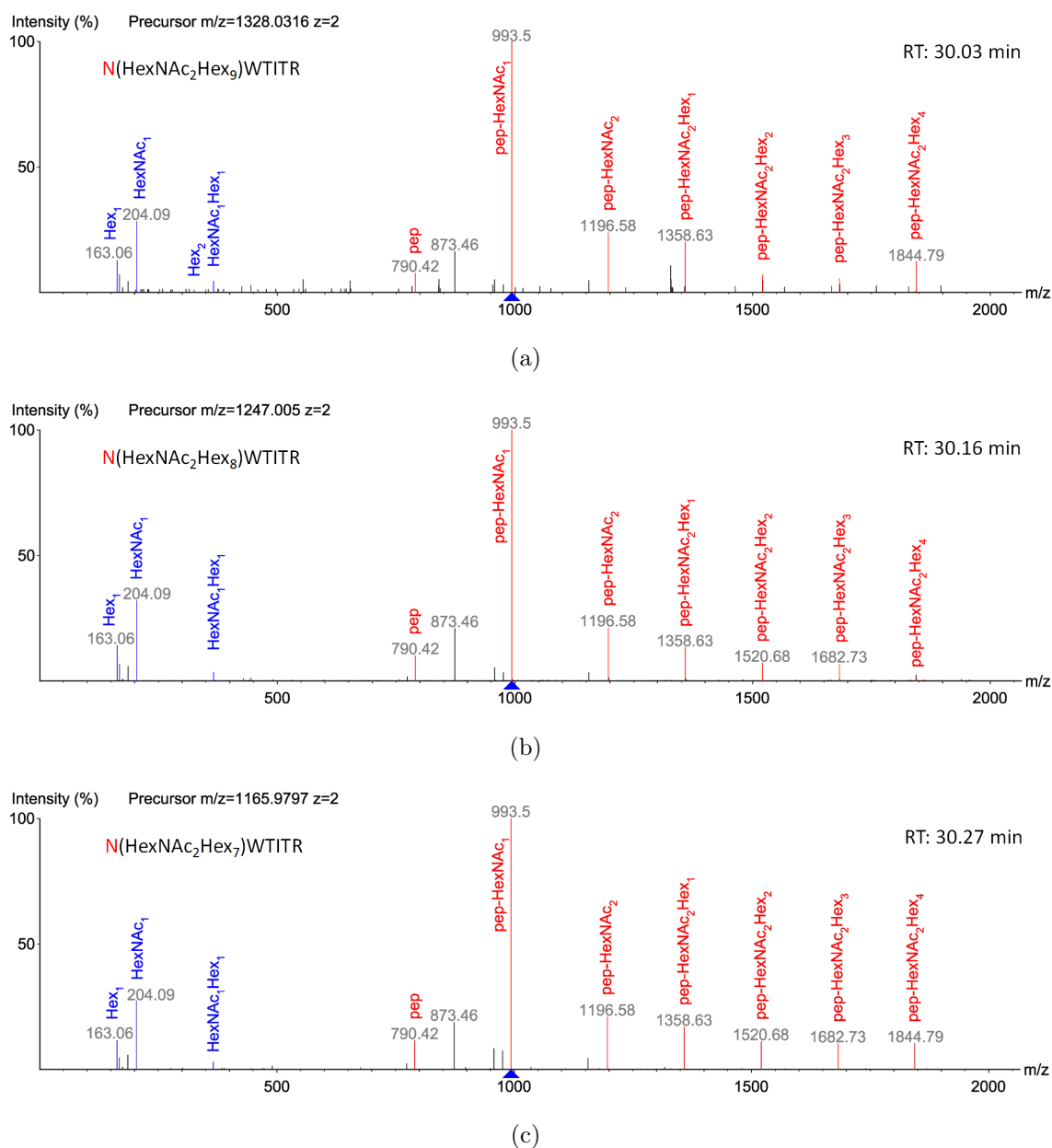


Figure 6.4: An example of glycans identified from three HCD spectra by GlycoMaster DB in the Enriched-HUP data set. Three HCD spectra have similar retention time but different precursor mass values. GlycoMaster DB identified three glycans. The calculated peptide mass is approximate 771.41. NWTITR is the only tryptic glycopeptide matching this mass value from the protein short list provided to GlycoMaster DB. The mass errors of the identifications of these three spectra are -1.29 ppm, -1.08 ppm, and -0.84 ppm, respectively.

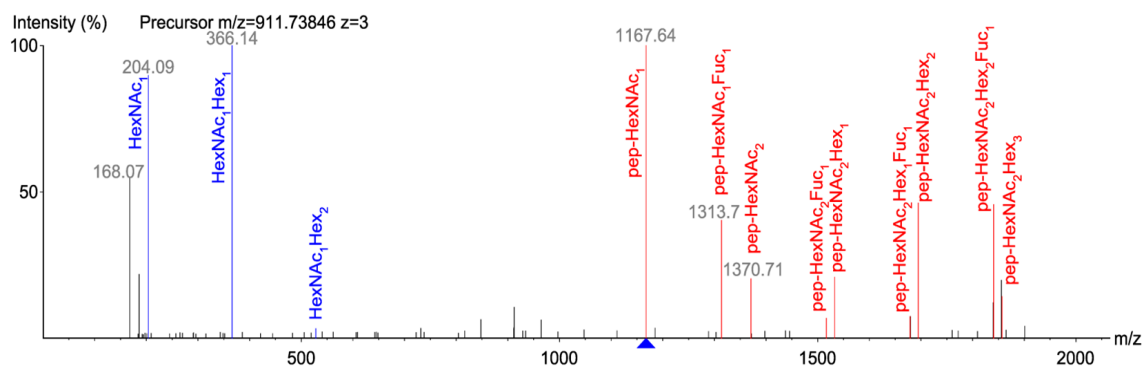
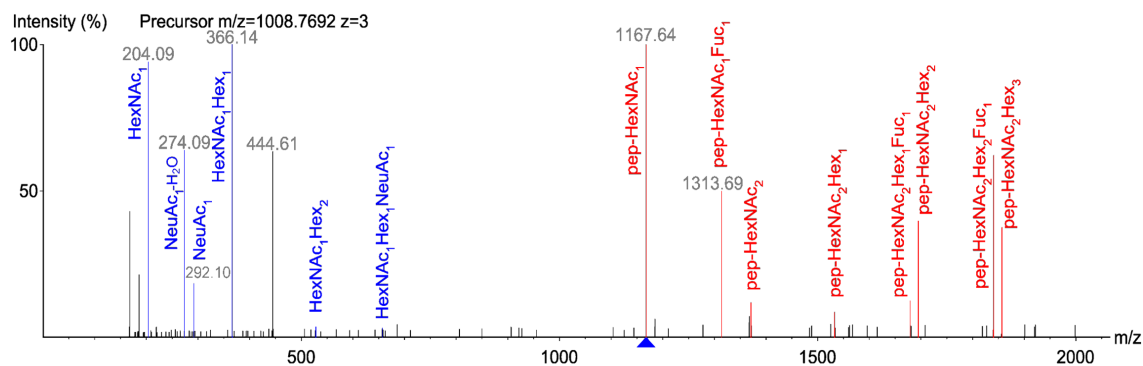

 (a) SLHVPGLN(HexNAc₄Hex₅Fuc₁)K, RT = 34.19 min

 (b) SLHVPGLN(HexNAc₄Hex₅Fuc₁NeuAc₁)K, RT = 36.89 min

Figure 6.5: Illustration of two HCD mass spectra that are interpreted as the same peptide but two slightly different glycans in the Enriched-HUP data set. (a) The oxonium ions from sialic acids are not present, and this indicates the absence of sialic acids in the glycan; (b) The peaks at m/z 292.10 and 274.09 indicate the existence of oxonium ions of sialic acid residues.

The number of identified proteins from each spectral file ranges between 61 and 362. A spectrum was analyzed using GlycoMaster DB if it was not identified either by PEAKS DB with $FDR \leq 1\%$ or by PEAKS *de novo* sequencing with $ALC \geq 50\%$. The results of those 30 spectral data were listed in Table 6.4. 337 spectra have matched glycans with high confidence ($-10 \lg P \geq 15$), and 298 of them have found corresponding peptide sequences from 229 proteins.

Figure 6.6 illustrates three example glycans identified by GlycoMaster DB from this data set.

6.2.5 Comparison of Identified Glycans between Enriched-HUP and HUP Data Sets

GlycoMaster DB identified 240 and 337 GSMs from Enriched-HUP and HUP data sets, respectively. Since the peptide sequences were searched only according to the calculated mass values, there might be ambiguities in the sequence identification. Moreover, the best matching structure from GlycoMaster DB might not be the real one because a spectrum could be matched equally well by several glycan structures sharing the same composition. Thus, the identified glycopeptides were grouped according to the combination of glycan composition and the peptide mass for each human urinary proteome data set. These groups, instead of individual glycopeptides, were compared to analyze the relationship of identified glycans between the two data sets. 141 and 201 such glycopeptide groups were discovered from the Enriched-HUP and HUP data set, respectively. Figure 6.7 is the Venn diagram that illustrates the overlaps between these two sets of glycopeptide groups.

The comparison reveals that GlycoMaster DB can identify more glycopeptides from the non-enriched data sets. The exact reason for this is unclear. But one probable reason is the possible different instrument settings in the generation of the HUP and Enriched-HUP data sets. In Enriched-HUP data set, a large number of MS/MS spectra are from the repeated fragmentation of the same precursor ion. Consequently, the total number of MS/MS scans is reduced from 199,890 to 33,954 after spectral merging, or a 6-fold reduction. While in the HUP data set, the phenomenon of repeated fragmentation of the same precursor is less severe. After the spectral merging, the reduction is from 170,215 to 72,152, or 2.4-fold.

CHAPTER 6. N-LINKED GLYCOPEPTIDE IDENTIFICATION

Table 6.4: This table lists the analysis result of each spectral data in HUP data set. Only the 23 spectral data having identified glycans by GlycoMaster DB are listed. The first column lists the names of the spectral files. The second column denotes the numbers of MS/MS spectra in each file after preprocessing. The third and fourth column lists the numbers of proteins reported by PEAKS DB and the numbers of un-interpreted spectra. The subsequent two columns give the number of spectra that passed the two filters, respectively. The number of identified glycans ($-10 \lg P \geq 15$) and peptides are listed in the last columns. The last row shows the total number of each column.

Data Name	MS/MS Number	Protein Number	Un-Interpreted MS/MS Number	Pass Filter-1	Pass Filter-2	Glycan Number	Peptide Number
ig06	891	71	241	35	6	2	2
ig07	928	61	348	69	18	6	3
ig08	878	70	235	43	3	2	2
ig09	811	87	306	49	5	4	4
ig12	892	78	209	35	12	8	8
ig13	868	79	182	52	15	11	6
ig14	4,222	274	931	339	80	32	28
ig15	4,553	340	1,093	447	108	47	42
ig16	4,146	289	1,017	335	56	32	31
ig17	3,924	312	953	321	47	23	22
ig18	3,822	325	1,123	390	45	24	21
ig19	3,717	309	1,087	335	38	16	15
ig20	4,152	325	1,172	429	46	23	23
ig21	4,106	278	1,259	502	54	25	22
ig22	3,847	287	1,251	401	37	17	13
ig23	4,094	303	1,332	500	63	13	11
ig24	4,884	308	1,512	432	29	8	8
ig25	4,205	326	1,495	395	33	6	6
ig26	4,107	310	1,536	422	35	9	8
ig27	3,960	300	1,677	428	26	7	5
ig28	3,828	292	1,740	317	19	8	5
ig29	4,139	362	1,751	509	44	10	9
iga3	1,178	87	257	37	5	4	4
Total	72,152	1,534 ^a	22,707	6,822	824	337	298

^aThis is the total number of unique proteins, rather than the sum of protein numbers reported in each spectral data.

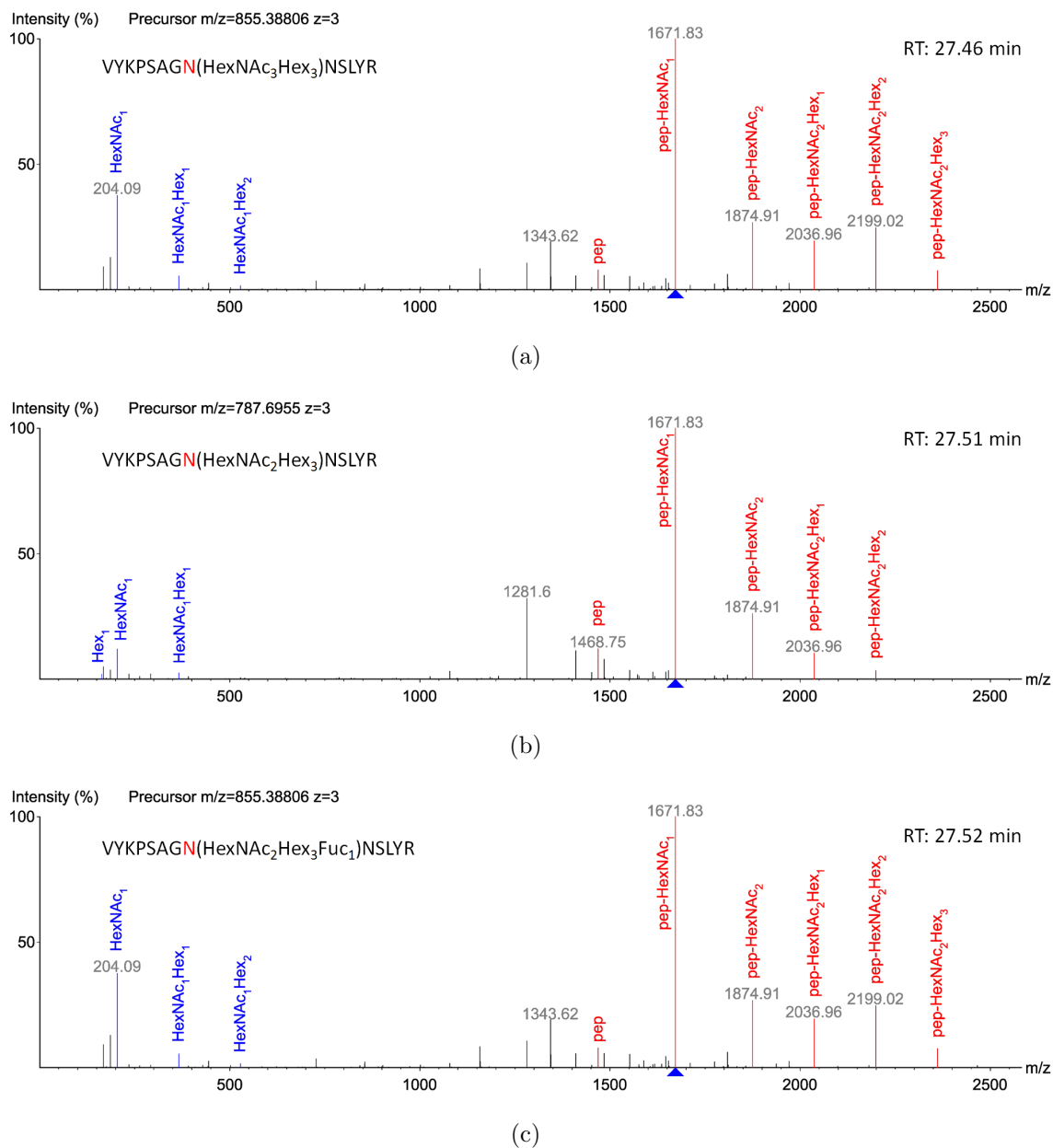


Figure 6.6: An example of glycans identified from three HCD spectra by GlycoMaster DB in the HUP data set. Three HCD spectra have similar retention time but different precursor mass values. GlycoMaster DB identified three glycans from them and these glycans differ from each other slightly. The calculated peptide mass is approximate 1449.74. VYKPSAGNNSLYR is one of the two peptides matching this mass value in the proteins provided to GlycoMaster DB but the other one has potassium adduct and much larger mass error at around 10 ppm. Therefore, VYKPSAGNNSLYR is selected as the glycopeptide and the precursor mass errors are 0.71 ppm, 0.08 ppm, and -0.51 ppm, respectively.

6.2.6 Glycopeptides with Same Mass

If only HCD data is available, the peptide is only reported according to the accurate mass and the presence of *N*-linked glycopeptide motif. This may result in ambiguous identification of the actual peptide when the size of the protein list is large or the mass accuracy is low. Computer simulation was carried out to study the severity of such ambiguity.

For each combination of mass accuracy (δ) and number (n) of proteins, n proteins were randomly selected from the UniProt human database (20,258 entries). The tryptic peptides containing the *N*-linked glycopeptide motif were generated *in silico*. The percentage of such peptides with unique mass (mass error $\leq \delta$ ppm) was calculated. The random selection was repeated 1,000 times for each δ and n , and the average percentage was plotted in Figure 6.8. It is noticeable that when the protein list contains no more than 100 entries, and the mass accuracy is better than 5 ppm, 99% of the tryptic peptides with the *N*-linked glycopeptide motif can be unambiguously identified from the mass.

6.3 Discussion

The experiments demonstrate the feasibility of using GlycoMaster DB to identify *N*-linked glycopeptides as well as the glycan structures (composition) from high-throughput HCD/ETD and HCD-only MS/MS data. The software is designed for the analysis of MS/MS data acquired from intact glycopeptides, rather than deglycosylated glycopeptides. Therefore, it can simultaneously identify glycans and peptide sequences. Such an application is important for large-scale glycoproteome analysis since the connection between glycans and their peptides can be readily determined. Figures 6.4, 6.5, and 6.6 illustrate multiple glycan forms on the same glycosylation sites. This is useful information to facilitate the study of the glycan synthesis and degradation process. In Figure 6.5, the different glycopeptides with the same peptide sequence have slightly different retention time. It excludes the possibility that the different forms are due to the post-source fragmentation in the mass spectrometer.

Most peaks in the HCD spectrum of a glycopeptide are from the fragmentation of the glycan but not the peptide. This makes it difficult to confidently identify the peptide sequence. Thus, a list of peptide sequences matching the calculated peptide

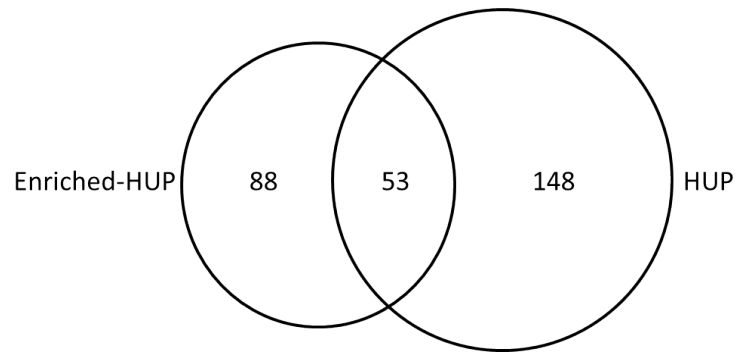


Figure 6.7: The Venn diagram showing the overlaps between the two sets of glycopeptide groups identified from Enriched-HUP and HUP data sets, respectively.

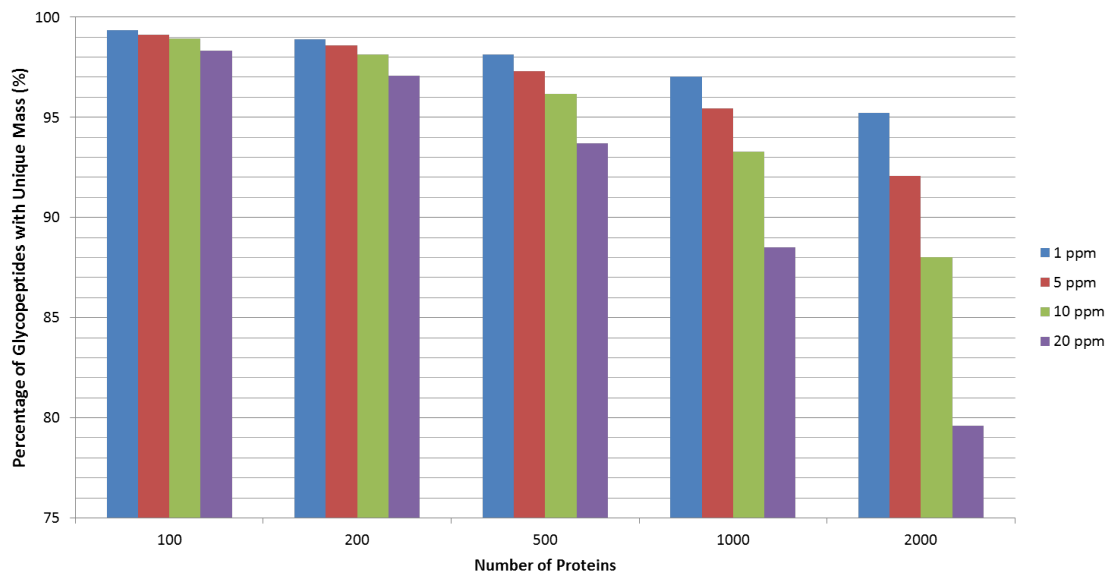


Figure 6.8: The average percentage of tryptic peptides containing the *N*-linked glycopeptide motif that have unique mass.

mass and containing the *N*-linked glycopeptide motif are reported. However, such identification may be ambiguous when there are a large number of proteins, especially when missed-cleavages and non-specific digestions were considered. The high mass accuracy of the LTQ-Orbitrap instrument can greatly help to determine the accurate precursor mass. In addition, the retention time of the glycopeptides is another piece of potentially useful information. However, the lack of a reliable retention time prediction algorithm for glycopeptides hinders the utilization of this information in our software. Future versions of GlycoMaster DB will consider including the retention time information when a glycopeptide retention time predictor becomes available. If the glycopeptides are fragmented with both HCD and ETD, GlycoMaster DB can use the spectrum-pairs simultaneously to identify both the glycans and the peptide sequences with high confidence.

Chapter 7

Maximum Peptide Feature Matching in Label-Free Quantification

The main difficulty of matching peptide features in a label-free quantification experiment is the inadequate reproducibility of the LC retention time. Due to factors such as aging, packing and contamination of an LC column, together with additional variability during experiment such as temperature, gradient shape and mixing physics, the retention time from different runs often shows large shifts and distortions. To match peptide features by using their mass and retention time information, the shifts and distortions need to be corrected. This is usually carried out by finding a monotonically increasing function $f(\cdot)$ that maps the retention time of a peptide in a sample to the retention time of the same peptide in the other sample. This process is often called the *retention time alignment*, or simply, *time alignment* [21, 86, 152].

It is noticeable that if the feature matching is available, the time alignment can be solved by fitting the times of the matched features with a smooth function. On the other hand, if the time alignment is known, the feature matching can be carried out by comparing the mass and the corrected time differences between features in the two samples. Although this is still not a trivial problem due to the existence of noise and false feature pairs, its solution is not dauntingly difficult. The real challenge of the feature matching problem lies in the mutual dependence between the time alignment function and the feature matching.

In the literature, the time alignment and the feature matching are usually dealt with in two separate steps. Some research in the literature uses heuristic algorithms to find an initial set of matched feature pairs, and then use these pairs to find a time alignment function. Conversely, some research find a time alignment first and then determine the matched feature pairs. Naturally, such procedures can be repeated iteratively to *hopefully* get a more and more accurate result.

This approach was typified by Li *et al.* [88], which matched features with similar m/z values as the initial feature matching. Kirchner *et al.* [81] used a robust point matching method to find an initial feature matching, and then carried out smooth monotone regression to find a time alignment. When there are significant time shifts and distortions, as well as the present of noisy false features, the finding of the initial set of feature matching in the above approaches can become challenging. However, this problem can be solved if the peptides of all the features are known since the features can be matched confidently by checking their peptide identities [44, 150]. This approach requires MS/MS spectra for the identification of the peptides. MS/MS analysis takes more duty cycles of the instrument. It reduces the number of MS scans so that many of the low-abundance peptides from the limited amount of biological samples become undetectable. Therefore, it is advantageous to perform quantification without MS/MS if a time alignment can be achieved without peptide identification. The peptides can be identified in a separate LC-MS/MS run after the quantification, possibly with an inclusion list that targets the quantified peptide features. In fact, there are even proposals in the literature to identify peptides purely based on the precursor m/z and the *aligned* retention time of a peptide feature [84]. This application definitely requires the time alignment without MS/MS. For these reasons, we assume the peptide identities are unknown to the alignment algorithm.

Other researchers focused on finding an initial time alignment function. Lange *et al.* [85] assumed that the time alignment is a linear function: $f(t) = a \cdot t + b$. A pair of coefficients (a_i, b_i) was calculated from every two pairs of possibly matched features. The correct coefficients (a, b) was estimated by finding a dense cluster of all the calculated (a_i, b_i) . However, the time alignment is actually nonlinear, so a number of publications [13, 20, 71, 109, 122, 121, 123, 136] only assume local linearity of the time alignment, and apply linear regression in small retention time intervals. These studies mostly differ at the methods used for (1) local linear regression, and (2) connecting the local linear regression results into a global time alignment function.

Although many of these methods have been used in practice, none of them defined a clear optimization goal for the peptide feature matching problem. There usually exist biological justifications for each step of these methods, but the property of the final output of a method is unclear. It is very different from the common practice in traditional algorithmic research, where the optimization goal is usually specified mathematically *before* the algorithm is being developed. Still, it is not uncommon in many emerging bioinformatics areas (including peptide quantification) that biologists often need a quick solution once an experimental method is invented. In such case an *ad hoc* solution is always useful. Moreover, the complexity of biology determines that the formulation of a tidy mathematical model is often difficult.

Disadvantages certainly exist in such *ad hoc* solutions. An immediate disadvantage is that the final outcome is unpredictable without running such an algorithm. The performance of the algorithm heavily depends on its implementation, such as the choice of parameters and the handling of some special cases. Therefore, a method developed in one lab has the tendency to get overfitting on its own data and may not work on the data from another lab or a new instrument. This recommended that a combinatorial problem should be clearly defined whenever it is possible. The separation of the problem formulation, the algorithm development, and the program implementation can help reduce the aforementioned overfitting tendency. The biological knowledge should be used exclusively during the problem formulation stage to specify the desired property of the solution. The algorithm development should strive to compute a solution that meets the specified property, instead of fitting the data that happen to be available for a researcher.

In this chapter, a clearly-defined combinatorial model for the feature matching problem is proposed in Section 7.1. The problem is proven to be *NP*-hard in Section 7.2. In Section 7.3, a slightly modified optimization goal is proposed, under which a polynomial time algorithm is presented. We show that the solution of the modified problem helps to determine an upper-bound and a lower-bound of the optimal solution. This results in a practical algorithm for the feature matching problem with a performance guarantee for each given instance. In Section 7.4, the optimization goal is amended to control the smoothness of the time alignment function for feature matching and a polynomial time algorithm is presented. Finally, Section 7.5 examined the performance of the algorithms on real LC-MS data. Not only is the proposed model tidy, but the performance of the algorithms also compares favorably with other existing methods.

7.1 The Maximum Feature Matching Problem

The peptide feature matching problem is formulated as a combinatorial optimization problem in this section. A *peptide feature* p is a 2-tuple $(m(p), t(p))$, where $m(p)$ indicates the mass and $t(p)$ indicates the retention time. We assume both $m(p)$ and $t(p)$ are integers since real numbers can be discretized by allowing a small rounding error. A *sample* consists of a set of features $\{p_1, p_2, \dots, p_n\}$. Let S and S' be two samples and their retention time ranges from 1 to T . A *time alignment function* that maps the time of S to the time of S' is a monotonically non-decreasing function $f : [1, T] \mapsto [1, T]$ such that $f(1) = 1$ and $f(T) = T$.

As aforementioned, the retention time of a peptide cannot be measured accurately. First, the unavoidable variations of LC conditions in the two MS runs can cause systematic drifts of the retention time for all peptides. This systematic error is modeled by the time alignment function f . Secondly, the retention time of an individual peptide may change independently from other peptides, causing a random error. Suppose two features $p \in S$ and $p' \in S'$ are from the same peptide, the random error is then modeled as $|t(p') - f(t(p))|$. After a proper time alignment, the random error is usually small. For example, if two LC runs are conducted on the same LC instrument under the same experimental condition and each lasts for an hour, the random error is often less than 1 minute after the time alignment.

For every two features $p \in S$ and $p' \in S'$, the matching quality of p and p' is measured by nonnegative function $w(\delta_m, \delta_t)$, where $\delta_m = |m(p') - m(p)|$ and $\delta_t = |t(p') - f(t(p))|$. The function w is also called a *weight function*. The *unit weight function*, denoted by w_I , is a straightforward definition of the weight function. Let $\Delta_m \geq 0$ and $\Delta_t \geq 0$ be the mass and time error tolerances, respectively. The unit weight function is then defined as

$$w_I(\delta_m, \delta_t) = \begin{cases} 1, & \text{if } \delta_m \leq \Delta_m \text{ and } \delta_t \leq \Delta_t, \\ 0, & \text{otherwise.} \end{cases} \quad (7.1)$$

The unit weight function treats a pair of features as a match if and only if their mass and time differences are within the error tolerances.

A *peptide feature matching*, or simply, a *feature matching*, is a bijective mapping between two subsets $P \subset S$ and $P' \subset S'$. More specifically, a feature matching provides a set of feature pairs, $M \subset \{(p, p') | p \in S, p' \in S'\}$, such that each feature

appears in at most one pair in M . Given a time alignment function f and a weight function w , the total weight of the matching M , is defined as

$$w(M) = \sum_{(p,p') \in M} w(|m(p') - m(p)|, |t(p') - f(t(p))|). \quad (7.2)$$

For label-free quantification, the two studied samples share most of their peptides and the biological experiments are optimized to minimize the noise and the mass and retention time errors. When the peptide identities for the peptide features are unknown, the most natural combinatorial goal for peptide feature matching is to maximize the total weight of the matching.

The *maximum feature matching problem* (MFM) is therefore defined as follows: Given two samples S and S' and a weight function w , find a time alignment function f and a feature matching M , such that $w(M)$ is maximized.

It can be noted that if f is given, MFM can be easily reduced to the maximum matching problem in a bipartite graph. In the reduction, each feature corresponds to a vertex and the two feature sets S and S' can be regarded as the two vertex sets of the graph. The weight of the edge between each feature pair is defined by the weight function w . In particular, when w is the unit weight function, the reduction results in the unweighted version of the maximum matching problem. It is well known that polynomial time algorithms exist for both weighted and unweighted maximum matching [46, 108]. However, for MFM, the time alignment function f needs to be computed simultaneously with the feature matching, and this makes MFM a much harder problem.

7.2 Maximum Feature Matching Is *NP*-Hard

Theorem 7.2.1 *The MFM problem is NP-hard under the unit weight function.*

Proof. The reduction is from the max-cut problem. Given an undirected graph $G = \langle V, E \rangle$, the max-cut problem splits the vertices into two disjoint sets V_1 and V_2 , such that $|\{(u, v) \in E \mid u \in V_1, v \in V_2\}|$ is maximized. It is well known that the max-cut problem is *NP*-complete [49].

Let $G = \langle V, E \rangle$ be an instance of the max-cut problem. Let $n = |V|$ and $m = |E|$.

For presentation clarity, a feature is visually shown as a data point on a mass-time grid (Figure 7.1). Each horizontal line on the plane corresponds to one time unit in the LC-MS experiment, and each vertical line corresponds to a mass unit. As MFM involves two samples S and S' , two colors, black and white, are used to distinguish the features in S and S' , respectively. A black feature is represented by a solid dot and a white feature is represented by a circle. It is possible that two features from S and S' have the same mass and time, and in such case the grid point is labeled with both a circle and a solid dot. Intuitively, MFM needs to match the black features onto the white features, allowing a small mass and time error after the time alignment. The time alignment can move all the black features on the same horizontal line up and down simultaneously, but it cannot change the relative time order of the black features.

The constructed instance of MFM consists of $T = 6n + 3$ time units and $6m$ mass units. Each edge $e_k \in E$ corresponds to a mass window of length 6, and each vertex v_i corresponds to a time window of length 6. The first and last time units are specially added and do not belong to any vertex time window. Figure 7.2 illustrates the construction that highlights edge e_k and vertex v_i . In the construction, many pairs of black and white features are added to certain grid points as shown in Figure 7.2. More precisely, a pair of black and white features are put at the grid point $(6k - 3, 6i - 4)$ for each $1 \leq k \leq m + 1$ and $1 \leq i \leq n + 1$. Additionally, for each edge e_k , two white features are put at $(6k - 4, 1)$ and $(6k - 4, T)$, respectively. The construction of the shaded areas in Figure 7.2 will depend on whether v_i is adjacent to e_k . Three cases arise: (1) $e_k = (v_i, v_j)$ and $i < j$; (2) $e_k = (v_j, v_i)$ and $i > j$; and (3) v_i is not adjacent to e_k . For each of the three cases, the construction of the shaded area is shown in Figure 7.3. Finally, we set the mass and time error tolerance as $\Delta_m = 1$ and $\Delta_t = 2$, and let the weight function be the unit weight function w_I . Thus, an instance of the MFM is constructed.

Within the mass window of an edge $e_k = (v_i, v_j)$, there are exactly two time windows (corresponding to v_i and v_j) that have the construction of Figure 7.3(a) and Figure 7.3(b), respectively. All other $(n - 2)$ time windows have the construction of Figure 7.3(c). Therefore, there are exactly $6n$ white features and $6n$ black features in the mass window. Consequently, the number of matches within a mass window is upper-bounded by $6n$.

If an identity time alignment function, $f(t) = t$, is used, the isolated black features

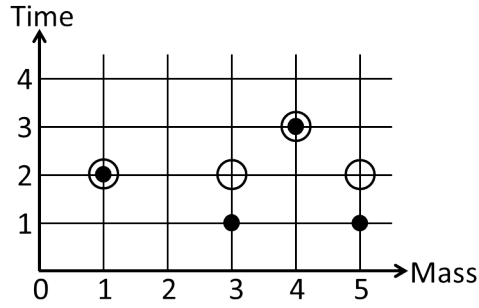


Figure 7.1: An Illustration of features plotting on a mass-time grid. Each horizontal line on the plane corresponds to one time unit in the LC-MS experiment, and each vertical line corresponds to a mass unit.

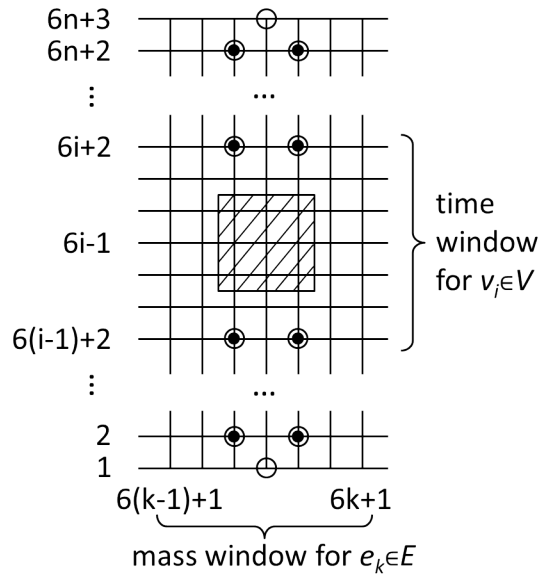


Figure 7.2: An illustration of the construction that highlights edge e_k and vertex v_i .

for v_i and v_j will not be matched. To match the isolated black feature of v_i in Figure 7.3(a), one can set either set $f(6i - 1) = 6i$ or $f(6i - 1) = 6i - 2$. Suppose the function $f(6i - 1) = 6i$ is selected. Then the black feature at time $(6i - 1)$ can be matched to the white feature at time $(6i + 2)$, because $|6i + 2 - f(6i - 1)| = 2 \leq \Delta_t$. Meanwhile, this will force the black feature at time $(6i + 2)$ to match the white feature at $(6i + 4)$, since the white feature at $(6i + 2)$ has already been matched by $f(6i - 1)$. This shifted matching propagates upward, as shown in Figure 7.4, until the isolated white feature at time T is matched. Similarly, if we let $f(6i - 1) = 6i - 2$ and match the isolated black feature of vertex i downward, then the shifted matching will propagate downward to use the white feature at time 1.

For each mass window for an edge $e = (v_i, v_j)$, there are only two isolated white features at time 1 and T , respectively. Therefore, the two isolated black features for v_i and v_j have to be matched to the opposite directions in order to be both matched, in which case the number of matches is exactly $6n$ for this mass window. However, if the two isolated black features are not matched to the opposite directions, only one of the two isolated white features can be used and the maximum number of matches is at most $6n - 1$.

Thus, if the max-cut has a solution $V = V_1 \cup V_2$ that cuts K edges, we can construct a time alignment function f , such that

$$f(t) = \begin{cases} t + 1, & \text{if } t = 6i - 1 \text{ and } v_i \in V_1, \\ t - 1, & \text{if } t = 6i - 1 \text{ and } v_i \in V_2, \\ t, & \text{otherwise.} \end{cases} \quad (7.3)$$

From the above discussion one can easily verify that each time window for a cut edge provides $6n$ matches and each of other time windows provides $6n - 1$ matches. The MFM instance has a total weight of $(6n - 1)m + K$.

On the other hand, suppose the MFM instance has $(6n - 1)m + K$ matches, and f is the time alignment function. Let $V_1 = \{v_i | f(6i - 1) \geq 6i\}$ and $V_2 = V \setminus V_1$. We get a solution for the max-cut instance. Because the number of matches in each mass window is upper bounded by $6n$, and there are $6(n - 1)m + K$ matches in total, we know that at least K mass windows provide $6n$ matches in each. For each of these K mass windows that corresponds to $e_k = (v_i, v_j)$, from the above discussion we know that the two isolated black features for v_i and v_j have to be matched to two opposite directions. In another word, the edge e_k is cut by separating its two vertices into V_1 and V_2 . Consequently, the constructed solution will cut at least K edges.

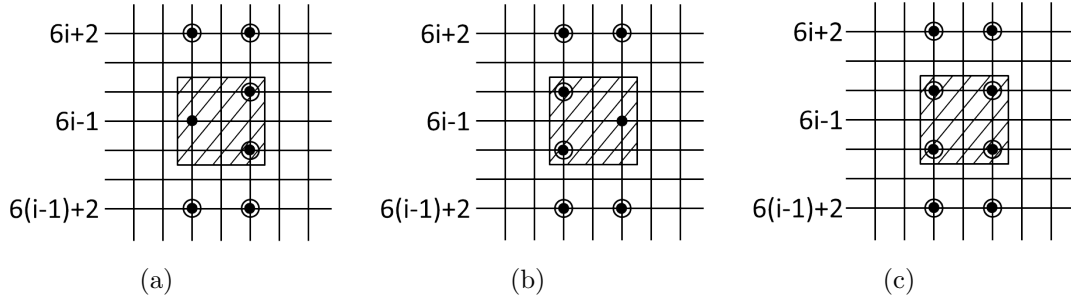


Figure 7.3: Illustrations of three cases to construct the grayed-out region: (a) $e_k = (v_i, v_j)$ and $i < j$; (b) $e_k = (v_j, v_i)$ and $i > j$; and (c) v_i is not adjacent to e_k .

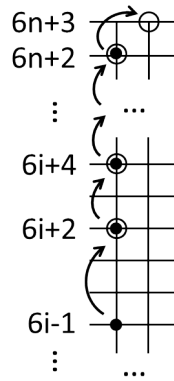


Figure 7.4: A shifted matching propagates upward when $f(6i - 1) = 6i$.

Thus, we have shown that $\text{max-cut} \leq_P \text{MFM}$, which proves the MFM problem is NP -hard. \square

7.3 A Practical Algorithm for Maximum Feature Matching

In this section we develop a practical algorithm for MFM. This is achieved by studying a slightly modified MFM problem. Instead of requiring the matching to be a bijective mapping, the modified problem only requires the matching to be a surjective mapping. More specifically, a surjective matching M^* is a subset of $\{(p, p') | p \in S, p' \in S'\}$, where $p \in S$ appears at most once and $p' \in S'$ can appear multiple times. Given a time alignment function f and a weight function w , the weight of the surjective matching M^* can be defined in the same way as in the MFM problem:

$$w(M^*) = \sum_{(p,p') \in M^*} w(|m(p') - m(p)|, |t(p') - f(t(p))|). \quad (7.4)$$

Given two samples and a weight function w , the *maximum surjective feature matching problem* (SFM) is to compute a time alignment function and a surjective matching M^* , such that $w(M^*)$ is maximized. We next present a polynomial time algorithm for the SFM problem.

For a sample S and a time i , let $S_i = \{p \in S | t(p) = i\}$ be the subset of features at time i . Let $S_{\leq i} = \{p \in S | t(p) \leq i\}$ be the subset of features with time at most i .

Let $d_{i,j}$ be the maximum weight of a surjective matching between S_i and S' that can be achieved by a time alignment function satisfying $f(i) = j$. Since the time of all features in S_i is equal to i and $f(i) = j$, $d_{i,j}$ can be easily computed by finding the best matching of each $p \in S_i$ separately.

Let $D_{i,j}$ be the maximum weight of a surjective matching between $S_{\leq i}$ and S' that can be achieved by a time alignment function satisfying $f(i) \leq j$. If $f(i) < j$, it is clear that $D_{i,j} = D_{i,j-1}$. If $f(i) = j$, the maximum surjective matching includes the maximum surjective matching from $S_{\leq i-1}$ to S' , and the maximum surjective matching from S_i to S' . Therefore, $D_{i,j} = D_{i-1,j} + d_{i,j}$. Combining the two cases, we know that $D_{i,j} = \max\{D_{i,j-1}, D_{i-1,j} + d_{i,j}\}$. With this recurrence relation, the SFM problem can be solved using a dynamic programming algorithm (Algorithm 3).

7.3. A PRACTICAL ALGORITHM FOR MAXIMUM FEATURE MATCHING

The optimal time alignment function f , as well as the surjective matching, can be computed by a standard backtracking.

Algorithm 3 Algorithm to solve the SFM problem.

- 1: **for** $i \leftarrow 0$ to T **do**
 - 2: **for** $j \leftarrow 1$ to T **do**
 - 3: Compute $d_{i,j}$
 - 4: **for** $i \leftarrow 0$ to T **do**
 - 5: $D_{i,0} \leftarrow 0, D_{0,i} \leftarrow 0$
 - 6: **for** $i \leftarrow 1$ to T **do**
 - 7: **for** $j \leftarrow 1$ to T **do**
 - 8: $D_{i,j} = \max\{D_{i,j-1}, D_{i-1,j} + d_{i,j}\}$
 - 9: Output $D_{T,T}$ as the maximum weight of the surjective matching.
-

Tracing back from $D_{T,T}$, all (i, j) pairs on the optimal path form the optimal time alignment function f .

Theorem 7.3.1 *The SFM problem can be solved in $O(T^2 + T \times |S| \times |S'|)$ time by Algorithm 3.*

Proof. The correctness of the algorithm is shown by the above discussion and the proof of the time complexity is as following. The computation of each $d_{i,j}$ in line 3 takes at most $O(|S_i| \times |S'|)$ time. Therefore, the whole loop at lines 1 to 3 takes time $O\left(\sum_{1 \leq i, j \leq T} |S_i| \times |S'|\right) = O(T \times |S| \times |S'|)$. After $d_{i,j}$ is computed and stored in memory, each execution of line 6 takes constant time. Thus, the loops from line 6 to line 8 take $O(|T|^2)$ time. \square

The computation of all $d_{i,j}$ is the most time-consuming part of Algorithm 3 and takes $O(T \times |S| \times |S'|)$. However, it is possible to speed up this part if the weight function w satisfies some properties.

Corollary 7.3.2 *If the unit weight function w_I is used, SFM can be solved in time*

$$O(T^2 + T \times |S| + |S| \times |S'|).$$

Proof. We only need to show that $d_{i,j}$ can be computed with time $O(T \times |S| + |S| \times |S'|)$ for all $1 \leq i \leq T$ and $1 \leq j \leq T$. For each $p \in S$, let $\mathcal{J}_p = \{j | \exists p' \in$

Algorithm 4 Algorithm to compute \mathcal{J}_p .

```

1:  $\mathcal{J}_p \leftarrow \emptyset$ 
2: for  $p' \in S'$  do
3:   if  $|m(p') - m(p)| \leq \Delta_m$  then
4:      $\mathcal{J}_p = \mathcal{J}_p \cup [t(p') - \Delta_t, t(p') + \Delta_t]$ 
    
```

S' such that $|m(p') - m(p)| \leq \Delta_m$ and $|t(p') - j| \leq \Delta_t$. \mathcal{J}_p can be computed by Algorithm 4.

In Algorithm 4 we need a data structure to store $\mathcal{J}_p \subset [1, T]$, which is the union of retention time intervals with the same length $2\Delta_t + 1$. Let A be a boolean array of length T that is used to store if $A[j]$ is the start position of one of the intervals. This structure can help to make the adding of a new interval take only $O(1)$ time.

To enumerate all $j \in \mathcal{J}_p$ takes at most $O(T)$ time, we propose Algorithm 5:

Algorithm 5 Algorithm to enumerate \mathcal{J}_p .

```

1: counter  $\leftarrow 0$ 
2: for  $j \leftarrow 1$  to  $T$  do
3:   if  $A[j]$  is true then
4:     counter  $\leftarrow 2\Delta_t + 1$ 
5:   if counter  $> 0$  then
6:     Output  $j$ 
7:     counter  $\leftarrow$  counter  $- 1$ 
    
```

Thus, the complexity of the Algorithm 4 is $O(|S'|)$. After \mathcal{J}_p is obtained, $d_{i,j}$ can be calculated by $d_{i,j} = |\{p \in S_i | j \in \mathcal{J}_p\}|$. The calculation of $d_{i,j}$ for all $1 \leq i \leq T$ and $1 \leq j \leq T$ can be carried out more efficiently with Algorithm 6. Since Algorithm 4 takes $O(|S'|)$ time for each $p \in S$, the accumulated time cost for line 2 is $O(|S| \times |S'|)$. Since $|\mathcal{J}_p| \leq T$, line 7 is repeated at most $O(\sum_{i=1}^T |S_i| \times T) = O(|S| \times T)$ times. Therefore, the total time complexity for Algorithm 6 is $O(T \times |S| + |S| \times |S'|)$. \square

Additionally, in Algorithm 4, if S' is sorted by mass values, we can retrieve all p' such that $|m(p') - m(p)| \leq \Delta_m$ by a binary search, without enumerating all $p' \in S'$. Because usually $\Delta_m \ll T$ and $|S'| > T$, this trick can significantly speed up in practice.

Algorithm 6 Algorithm to calculate $d_{i,j}$.

- 1: **for** $i \leftarrow 1$ to T **do**
 - 2: Calculate \mathcal{J}_p for each $p \in S_i$ with Algorithm 4;
 - 3: **for** $j \leftarrow 1$ to T **do**
 - 4: $d_{i,j} \leftarrow 0$
 - 5: **for** $p \in S_i$ **do**
 - 6: **for** $j \in \mathcal{J}_p$ **do**
 - 7: $d_{i,j} = d_{i,j} + 1$
-

Lemma 7.3.3 *Suppose two instances of SFM and MFM share the same input, the weight of the maximum feature matching (MFM) is less than or equal to the weight of the maximum surjective feature matching (SFM).*

Proof. A bijective mapping is also surjective. This indicates that a solution to MFM is also a solution to SFM. \square

There exists a straightforward way to convert the optimal solution for SFM to a suboptimal solution for MFM. Let $M^* \subset \{(p, p') | p \in S, p' \in S'\}$ be a solution to SFM, such a conversion can be done by selecting only one pair of features from M^* for every $p' \in S'$. Furthermore, Algorithm 7 can generate a better suboptimal solution for MFM based on the optimal solution to SFM.

Algorithm 7 Algorithm SMFM to provide a suboptimal solution for the MFM problem.

- 1: Compute an optimal solution for SFM using Algorithm 3;
 - 2: Let f be the optimal time alignment function;
 - 3: Let w_u be the optimal weight of SFM;
 - 4: **for** $p \in S$ **do**
 - 5: **for** $p' \in S'$ **do**
 - 6: $\tilde{w}(p, p') = w (|m(p') - m(p)|, |t(p') - f(t(p))|)$
 - 7: Treat $\tilde{w}(p, p')$ as the edge weight in a complete bipartite graph $S \times S'$, and compute a maximum bipartite matching;
 - 8: Let w_l be the weight of the maximum bipartite matching;
 - 9: Output the maximum bipartite matching as the suboptimal solution to MFM, w_u as the upper bound of the optimal weight of MFM, and w_l as the lower bound.
-

Theorem 7.3.4 *Algorithm SMFM computes a suboptimal solution for the MFM problem, and determines an upper-bound and a lower-bound of the optimal weight.*

Proof. The theorem is an immediate consequence of Lemma 7.3.3. □

Since MFM is *NP*-hard, there is no polynomial time algorithm for finding the optimal solution unless $P=NP$. Therefore, Algorithm SMFM can be used in practice to find a suboptimal solution.

7.4 Variations of the Maximum Feature Matching Problem

In this section we examine two variations of the MFM problem. A more accurate weight function is introduced in the first variation, and a gap penalty for the time alignment is added in the second variation. The gap penalty can help make the time alignment function smoother.

7.4.1 Weight Function

The unit weight function w_I is conceptually simple and the mass and time error thresholds Δ_m and Δ_t can be easily determined by the technician who operates the instrument according to experience. However, it is sometimes desirable to use a continuous weight function to give different weights to different time errors.

It has been shown that in real data the random retention time error after the time alignment satisfies a normal distribution [43]. Let $\epsilon_i = t(p'_i) - f(t(p_i))$ be the random time error of a pair of matched features (p_i, p'_i) after the time alignment and $\Pr(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\epsilon_i}{\sigma}\right)^2}$ be the probability distribution of ϵ_i . Assume the random error of different features are independent to each other, then the probability of all the errors in the matching is $\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\epsilon_i}{\sigma}\right)^2}$. By taking the logarithm, it is easy to see that maximizing the above probability is equivalent to minimizing $\sum_{i=1}^n \epsilon_i^2$. Because the weight function needs to be nonnegative, we define the following weight function w_2 :

$$w_2(\delta_m, \delta_t) = \begin{cases} \Delta_t^2 - \delta_t^2, & \text{if } \delta_m \leq \Delta_m \text{ and } \delta_t \leq \Delta_t, \\ 0, & \text{otherwise.} \end{cases}$$

7.4.2 Gap Penalty

In the definition of the MFM and SFM problems, the time alignment function f is only restricted to be a monotonically increasing function. However, it is sometimes beneficial to require some smoothness of f since fewer data points are required to fit a smooth function.

Let $[l_i, r_i]$ ($i = 1, \dots, k$) be the maximal time intervals such that $r_i - l_i > 1$ and $f(t)$ remains a constant in each interval. These are called the type-I gaps. The gap length for $[l_i, r_i]$ is $r_i - l_i$. Let $[l'_i, r'_i]$ ($i = 1, \dots, k'$) be the maximal time intervals such that there is no t satisfying $f(t) \in [l'_i, r'_i]$. These are called the type-II gaps. The gap length for $[l'_i, r'_i]$ is $l'_i - r'_i + 1$. By requiring f to be smooth, we essentially want to penalize these two types of gaps with a gap penalty function $g(k) > 0$ for a length- k gap.

This is analogous to the gaps in the pair-wise sequence alignment. The major difference is that here we prefer many smaller gaps over a few large gaps. Therefore, in contrast to using a concave gap penalty function in a sequence alignment, a convex gap penalty function, such as $g(k) = k^2$, is chosen, and the total gap penalty of the time alignment function f is defined as

$$g(f) = \sum_{i=1}^k g(r_i - l_i) + \sum_{i=1}^{k'} g(r'_i - l'_i + 1). \quad (7.5)$$

The *gapped-MFM* problem is to find a bijective feature matching M and a time alignment f to maximize $\text{score}(M, f) = w(M) - g(f)$. Similarly, the *gapped-SFM* problem is to find a surjective feature matching M^* and a time alignment f to maximize $\text{score}(M^*, f) = w(M^*) - g(f)$.

We design a dynamic programming algorithm for the gapped-SFM problem. Let $K > 0$ be the maximum allowed gap length. Let S_i , $S_{\leq i}$ and $d_{i,j}$ be as defined in Section 7.3. Let $N_{i,j}$ be the maximum score achieved by features in $S_{\leq i}$ and a time alignment function satisfying $f(i) = j$ and $f(i-1) < j$. Let $M_{i,j}$ be the maximum score achieved by features in $S_{\leq i}$ and a time alignment function satisfying $f(i) = j$.

From the definition of $N_{i,j}$, we know that $[f(i-1)+1, f(i)-1]$ is a probable type-II gap. Let $k = f(i) - f(i-1) - 1$ be the gap length. Then, $f(i-1) = f(i) - k - 1 = j - k - 1$, and therefore

$$N_{i,j} = \max_{0 \leq k \leq K} \{M_{i-1, j-k-1} + d_{i,j} - g(k)\} \quad (7.6)$$

To compute $M_{i,j}$, assume that $i - k$ is the least number such that $f(i - k) = j$. Then $[i - k, i]$ is a probable type-I gap, and therefore,

$$M_{i,j} = \max_{0 \leq k \leq K} \{N_{i-k,j} + \sum_{l=i-k+1}^i d_{l,j} - g(k)\} \quad (7.7)$$

From Eq. (7.6) and Eq. (7.7), it is straightforward to develop a dynamic programming algorithm (Algorithm 8) to compute $N_{i,j}$ and $M_{i,j}$ simultaneously. The time complexity will be $O(T^2K)$ plus the time needed by computing $d_{i,j}$. Therefore, for a general weight function w , the time complexity is $O(T^2K + T \times |S| \times |S'|)$.

Algorithm 8 Algorithm SFM-g to solve the gapped-SFM problem.

```

1: for  $i \leftarrow 0$  to  $T$  do
2:   for  $j \leftarrow 1$  to  $T$  do
3:     Compute  $d_{i,j}$ 
4:   for  $k \leftarrow 0$  to  $T$  do
5:      $N_{0,k} \leftarrow -g(k)$ 
6:      $M_{k,0} \leftarrow -g(k)$ 
7:   for  $i \leftarrow 0$  to  $T$  do
8:     for  $j \leftarrow 1$  to  $T$  do
9:        $N_{i,j} = \max_{0 \leq k \leq K} \{M_{i-1,j-k-1} + d_{i,j} - g(k)\}$ 
10:       $M_{i,j} = \max_{0 \leq k \leq K} \{N_{i-k,j} + \sum_{l=i-k+1}^i d_{l,j} - g(k)\}$ 
11: Output  $N_{T,T}$  as the maximum weight of the surjective matching.
```

Since MFM is *NP*-hard, gapped-MFM with a general gap penalty is also *NP*-hard. Algorithm SMFM in Section 7.3 can be slightly modified to Algorithm SMFM-g to provide a suboptimal solution for gapped-MFM and an upper bound to the optimal score. The only required modification is to use Algorithm SFM-g in line 1 of the algorithm, instead of using the algorithm for SFM (Algorithm 3).

7.5 Experimental Results

The performance of our algorithms was compared with three other state-of-the-art software tools, msInspect [13], MZmine2 [121], and MultiAlign [84] by using real LC-MS data sets. Our algorithms include: (1) Algorithm SMFM with the weight function

w_2 , and (2) the algorithm with weight function w_2 and a gap penalty $g(k) = 10k^2$, as described in Section 7.4. For the rest of the section, the first algorithm will be denoted by *SMFM*, and the second algorithm will be denoted by *SMFM-g*.

Five LC-MS data sets, which were produced from the yeast proteome by three different labs, were chosen for the performance evaluation. All of these data sets were published in previous research [7, 142, 110]:

iPRG2011: The data was from the “Proteome Informatics Research Group study” in 2011 [7]. *Saccharomyces cerevisiae* lysate is digested by Lys-C followed by strong cation exchange chromatography (SCX) fraction. 10 fractions were selected from 15 fractions. Each fraction is was analyzed by LC-MS/MS with a Thermo LTQ-Orbitrap XL. The LC separation is done in a flow rate of 200nl/min using a $75\mu\text{m} \times 10$ cm column packed with 3um particle size Repronil C18AQ (Solvent A: 0.1% formic acid, Solvent B: 90% acetonitrile/0.1% formic acid, Gradient: load at 3%B, elute with 5-35%B in 90 min, elute with 35-90%B in 10 min; wash at 90%B for 9 min. 10-20 sec chromatographic peak widths). Orbitrap was used to collect high resolution MS spectra. 8 most abundant precursor were fragmented in data dependent mode to produce MS/MS. There are 22,087 MS scan and 103,185 MS/MS scan for all these 10 fractions. In our experiment the LC-MS data of fraction 1 was picked from the 10 published fractions for software performance evaluation.

Coon: This data was produced from two biological replicates in the Coon research group [142]. A whole cell yeast lysate was digested using the protease endo-LysC and was separated into 12 fractions by SCX fraction. Each fraction is loaded in online nanoflow reversed-phase liquid chromatography coupled to MS/MS (nLC-MS/MS), using a forty minute linear gradient of 1.4% to 49% acetonitrile in 0.2% formic acid with data-dependent precursor selection. Eluting peptide cation populations were analyzed using the Orbitrap for MS and QLT for MS/MS product ion spectra. We chose two LC-MS data sets from fraction 3 of biological replicate 1 (*Coon1.F3*) and fraction 4 of biological replicate 2 (*Coon2.F4*) in our experiment. *Coon2.F4* shared the most number of peptides with the iPRG data set, and *Coon1.F3* was the fraction in replicate 1 that shared the most number of peptides with *Coon2.F4*.

Mann: The data was from the single-shot LC-MS/MS system in Mann Lab [110] measured six yeast cell lysate separately. Each is digested by LysC digestion using the FASP method. Peptides were loaded on a 50 cm column with $75\text{-}\mu\text{m}$ inner diameter, packed in-house with $1.8\text{-}\mu\text{m}$ C_{18} particles (Dr. Maisch GmbH, Germany). Reversed

Table 7.1: The number of features in different samples.

	iPRG	Coon1.F3	Coon2.F4	Mann.1	Mann.2
Feature Number	11,430	5,879	5,320	66,479	68,128

phase chromatography was performed using the Thermo EASY-nLC 1000 with a binary buffer system consisting of 0.5% acetic acid (buffer A) and 80% acetonitrile in 0.5% acetic acid (buffer B). The peptides were separated by a linear gradient of buffer B up to 40% in 240 min for a 4h gradient run with a flow rate of 250 nl/min in the EASY-nLC 1000 system. Eluting peptides were analyzed on the bench-top quadrupole Orbitrap mass spectrometer(Q Exactive) and the top 10 abundant peptide ions in a survey scan were fragmented using HCD. We chose the first two biological replicates (*Mann.1* and *Mann.2*) for the software performance evaluation in our experiment.

The names of the data sets and the number of features detected by msInspect in each of them are listed in Table 7.1. These five data sets are aligned with one another under different settings. More specifically, the alignments **Coon1.F3 vs. Coon2.F4** and **Mann.1 vs. Mann.2** are data sets from the same lab on the same instrument in the same experiment. These reflect the easiest test cases since the LC conditions do not vary too much. The alignments **iPRG vs. Coon2.F4** and **Coon2.F4 vs. Mann.1** reflect the most challenging test cases, since the aligned data sets were from different labs and the LC conditions across different labs present the largest possible variations. However, since they were all produced from the yeast proteome, there should be a significant number of peptides shared by the data sets. Therefore a robust feature matching algorithm should still be able to match these common peptides' features, despite the existence of large retention time distortion and noises.

For each data set, the MS/MS spectra were used to identify peptides with the PEAKS [166]. The yeast protein database was searched with following search parameters:

- parent mass error tolerance = 20 ppm (part-per-million);
- fragment mass error tolerance = 0.5 Da;

- fixed PTM: carbamidomethylation on Cys (+57.02);
- variable PTMs: deamidation on Asn and Gln (+0.98), methyl ester on Lys (+14.02), and oxidation on Met (+15.99).

The peptides identified with $\text{FDR} \leq 1\%$ and matched by only one feature in the LC-MS data were selected as a control set. This control set was a subset of “true” peptide feature matches between different data sets and used to evaluate different software’s performance.

Each of the compared software tools, SMFM, SMFM-g, msInspect, MZmine2, and MultiAlign, was used to produce the pairwise time alignment for iPRG vs. Coon2.F4, Coon2.F4 vs. Mann.1, Coon1.F3 vs. Coon2.F4, and Mann.1 vs. Mann.2, respectively. The m/z and retention time error tolerance of each software were set to be the same whenever possible. More specifically, Δ_t was set to be five minutes for the samples from different labs (iPRG vs. Coon2.F4 and Coon2.F4 vs. Mann.1) and two minutes for the ones from the same lab (Coon1.F3 vs. Coon2.F4 and Mann.1 vs. Mann.2). Other unique parameters of a software tool were set separately to achieve its own best performance:

1. SMFM: $\Delta_m = 15$ ppm (part-per-million).
2. SMFM-g: $\Delta_m = 15$ ppm, gap penalty $g(k) = 10k^2$.
3. msInspect: spline mode, mass error tolerance = 15ppm.
4. MZmine2: RANSAC algorithm mode, m/z tolerance = 10 ppm¹, retention time tolerance (before correction) = 50 minutes, number of RANSAC iterations = auto, minimal number of points = 20%, threshold value = 3, and same charge state was required.
5. MultiAlign: mass tolerance = 15 ppm, and hybrid recalibration was selected.

The peptide features detected by msInspect from the LC-MS raw data were exported as the input of SMFM, SMFM-g, and msInspect. MultiAlign and MZmine2 do not accept features detected by msInspect. Therefore, MultiAlign used the features detected by DeconTools [138] which was the preferred feature detection method of MultiAlign. MZmine2 used its own feature detection result.

¹Error tolerance 15ppm crashed the software

Table 7.2: The comparison of average aligned time errors (in seconds) and the percentages of correctly aligned feature pairs on true peptide features.

	SMFM	SMFM-g	msInspect	MZmine2	MultiAlign	Polynomial-4
iPRG-	36.6	35.2	114.8	55.0	126.0	30.3
Coon2.F4	(100%)	(100%)	(87%)	(99%)	(92%)	(100%)
Coon2.F4-	63.9	62.1	97.1	65.7	78.7	66.2
Mann.1	(82%)	(82%)	(71%)	(80%)	(73%)	(79%)
Coon1.F3-	8.4	7.4	11.3	21.8	27.8	8.2
Coon2.F4	(100%)	(100%)	(96%)	(94%)	(89%)	(100%)
Mann.1-	14.5	13.0	-	-	16.4	15.7
Mann.2	(90%)	(87%)	-	-	(81%)	(86%)

The performance of each method was measured quantitatively with the average aligned time error of the true feature pairs. More specifically, for each pair of features $p = (m(p), t(p))$ and $p' = (m(p'), t(p'))$ that were from the two compared samples and shared the same peptide, the aligned time error was calculated as $|f(t(p)) - t(p')|$, where $f(\cdot)$ was the retention time alignment function calculated by each software tool. The average aligned time error and the percentage of correctly aligned “true” feature pairs of each software applying on each pair of data sets are provided in Table 7.2. A feature pair is considered as correctly aligned if their aligned retention time difference is below the specified threshold in each experiment.

Although the five above mentioned software tools did not use the peptide identification deliberately, just for curiosity, the average aligned time errors obtained by a simple method (Polynomial-4) that used the peptide identification were also added in Table 7.2. By using the true feature pairs derived from the peptide identification, the Polynomial-4 method fitted a fourth degree polynomial as the time alignment function.² This was an unfair comparison because Polynomial-4 used additional information. Nevertheless, Table 7.2 showed that our new methods SMFM and SMFM-g also compared well to this polynomial fitting. This indicated that the time alignment function could not be fit accurately by a low degree polynomial, and further justified the use of a monotonically increasing function instead of any specific simple function in our SMFM model. For the alignment of Mann.1 vs. Mann.2, both msInspect and

²The second and third degree were also tried but the results were not as good as the fourth degree.

MZmine2 failed (msInspect crashed and MZmine2 returned no result). We suspected that it was due to the large data size of Mann’s data sets (see Table 7.1). Our new algorithms (SMFM and SMFM-g) finished successfully in less than one minute with 560 MB of memory usage.

Figure 7.5 illustrates the relative performance of the six compared methods visually. The resulting time alignment from each software was plotted together with the “true” peptide feature pairs (represented by blue circles). Retention time of both samples were scaled to 3,600 seconds in the figure. All the possible feature pairs that had a mass difference less than 15 ppm were also plotted as gray crosses. Thus, intuitively, the software tools were using these gray crosses to compute a time alignment function. A better software tool can generate an alignment function that fits the trend of blue circles.

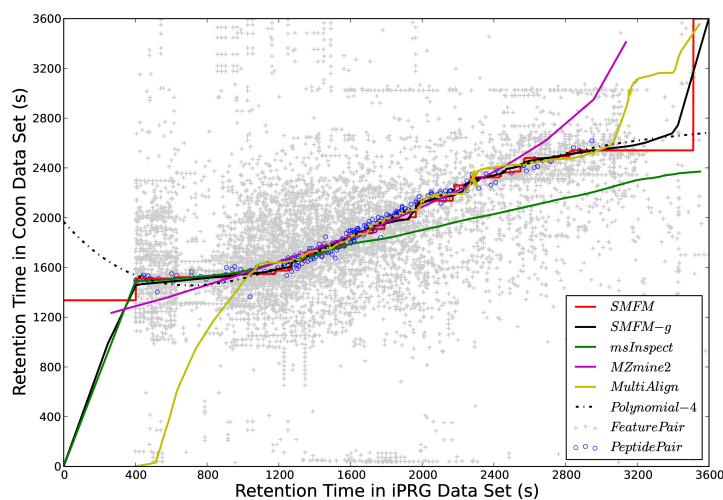
Similar figures for the alignments between biological replicates, Coon1.F3 vs. Coon2.F4 and Mann.1 vs. Mann.2, were plotted in Figure 7.6. The time alignment functions on these data sets were almost linear functions.

7.6 Discussion

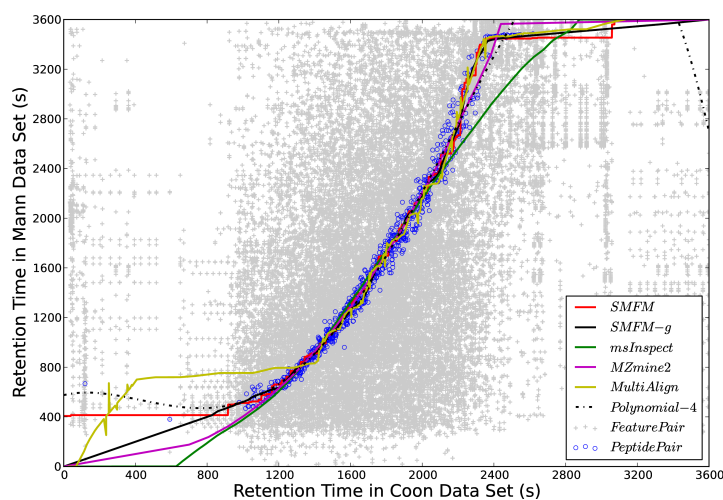
The maximum feature matching problem (MFM) is formulated to match the peptide features in label-free peptide quantification. To our knowledge this is the first combinatorial model for the problem. We show that the problem is *NP*-hard and provide practical algorithms that guarantee the performance for each instance. Experiments on real data demonstrate that our algorithms have better performances comparing to other software in the literature.

While recognizing the requirement and contribution of *ad hoc* software tools in bioinformatics research, we advocate that, whenever possible, a bioinformatics problem should have a clear combinatorial definition. This traditional practice in algorithmic research can help reduce the risk of overfitting the training data in the process of seeking for a better algorithm. It also helps predict the performance of an algorithm before implementing and running the software.

A feature p is defined by a pair $(m(p), t(p))$ in our study. However, more information about a peptide feature retrieved from the LC-MS data can be added by replacing $m(p)$ with an information vector. Meanwhile, in the wight function $w(\delta_m, \delta_t)$, δ_m

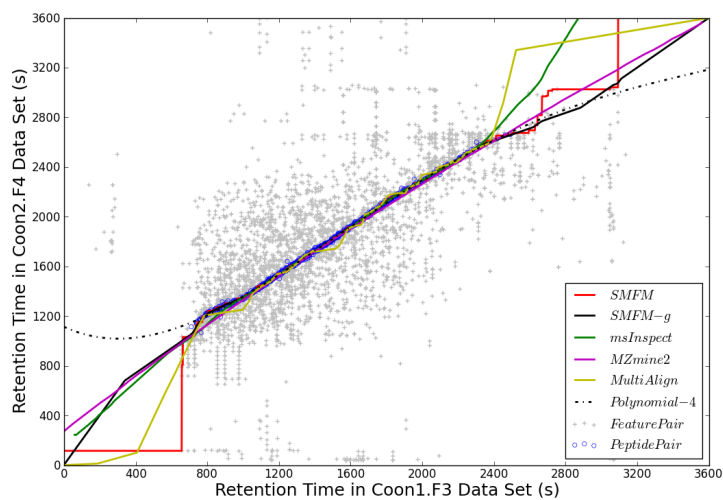


(a) iPRG vs. Coon2.F4

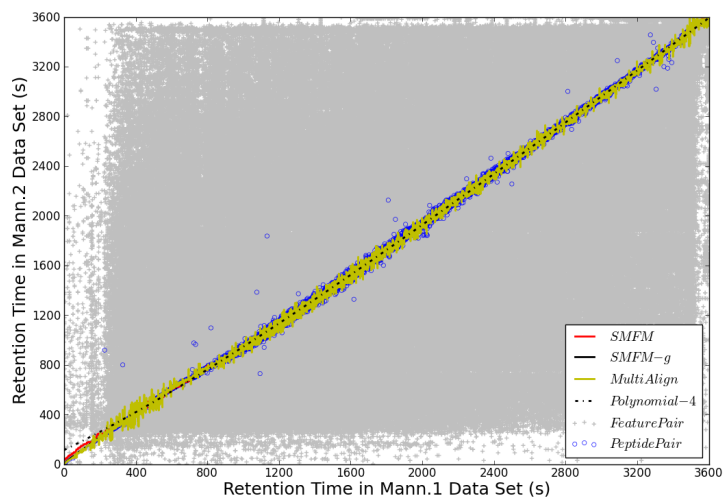


(b) Coon1.F3 vs. Mann.1

Figure 7.5: Comparison of the feature matching software tools on data sets from different labs: iPRG vs. Coon2.F4 (a) and Coon1.F3 vs. Mann.1 (b). The x-axis denotes the retention time in the first sample and the y-axis denotes the retention time in the second sample. A blue circle, which stands for a feature pair matched according to peptide identification, is considered as the ground truth. A gray cross represents a possible feature pair matched purely by the precursor mass. The curves are produced by the compared algorithms without knowing the blue circles.



(a) Coon1.F3 vs. Coon2.F4



(b) Mann.1 vs. Mann.2

Figure 7.6: Comparison of the feature matching software tools on data sets from the same lab: Coon1.F3 vs. Coon2.F4 (a) and Mann.1 vs. Mann.2 (b). As msInspect and MZmine2 failed to align the data set Mann.1 and Mann.2, only results of SMFM, SMFM-g, MultiAlign, and Polynomial-4 are shown in (b).

needs to be replaced by the distance of the two vectors of the compared features. For example, the intensity distribution over the isotopic peaks and over the retention time can be used to measure the similarity (or matching quality) of two matched features. The *NP*-hardness and algorithms remain the same in such case.

Researchers develop bioinformatics software to help find “real” biological solutions from their experimental data. However, as the real solution is unknown before using the software, the optimization goal is at most an approximation to the *properties* of the real solution, instead of the real solution itself. We have demonstrated that a clear definition of such an optimization goal has converted a biological problem to a pure combinatorial problem that is readily for algorithmic research. Meanwhile, performances of the algorithms that proposed for this combinatorial problem compare favorably to the state-of-the-art *ad hoc* software packages.

In fact, a clear definition of the optimization goal is helpful even in *ad hoc* solutions. For example, previous research has suggested alternately finding a time alignment and a set of matching features by using each other as the input. It is not guaranteed that such iteration can converge or improve the result. However, if the optimization goal is definite, the iteration can be evaluated after each loop and terminated when a certain requirement is achieved.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this dissertation, MS/MS-based computational proteomics is explored on both protein identification and quantification. The major content focuses on peptide identification, especially the modified peptide identification.

Protein identification is to interpret the spectral data and thus retrieve the protein and PTM information. A large number of software packages, using database search or *de novo* sequencing approaches, have been developed and undoubtedly accelerated the progress of proteomics studies. However, with the rapid development in experimental strategies and the mass spectrometry instruments, people are not satisfied by current status in computational proteomics. On one hand, the accuracy and the resolution of the instruments are kept getting better, and it is more flexible to choose different techniques in an experiment. Thus, it is theoretically possible to improve the accuracy of *de novo* sequencing and make it become a more practical approach. On the other hand, only identifying proteins in a given sample is still far away from the goals of proteomics. A thorough study on proteins and their PTMs demands sophisticated, and specifically designed computational approaches. Thus, this dissertation provides algorithmic solutions to fulfil these urgent requirements.

MS/MS instruments that implement multiple fragmentation modules can generate different types of mass spectral data for the same sample. This inspired us to design a novel scoring scheme, ADEPTS, to improve the performance of *de novo* sequencing in Chapter 3. Features from two types of spectra are considered simultaneously in

this scoring scheme. The comparison between ADEPTS and other software tools, including PEAKS on CID and ETD data separately, PepNovo on CID data, and CompNovo on CID/ETD spectrum-pairs, shows that ADEPTS performs better on both correctly identified peptides and residues.

Novel proteins can be sequenced using *de novo* sequencing approaches, but identifying PTMs of these proteins is a nontrivial challenge. Conventionally, users need to specify many PTMs that possibly exist in a sample to a *de novo* sequencing software tool. It increases the running time and degrades the result accuracy significantly. Our novel dynamic programming algorithm, DeNovoPTM, is proposed in Chapter 4 to solve this problem by limiting the number of PTM occurrences per peptide. Experiments show that DeNovoPTM outperforms other two state-of-the-art *de novo* sequencing software, PEAKS and PepNovo, on modified peptide identification when many PTMs types are considered.

Database search has been regarded as a reliable protein identification approach. However, conventional database search tools cannot provide a PTM search efficiently and accurately when a large number of PTMs are specified. We propose PeaksPTM in Chapter 5 as an improved database search approach to enable the unrestricted PTM identification. Different from conventional database search tools, PeaksPTM does not require users to specify PTMs in advance; instead, all the PTMs recorded in the Unimod database are considered in the search by default. A modified target-decoy strategy is also applied to control false positives. PeaksPTM makes it possible to unrestrictedly and confidently identify the general PTMs existing in a complex biological sample. Experiments show that PeaksPTM achieves a stronger performance than competitive tools for unrestricted identification of PTMs.

Glycosylation is one of the most frequently observed PTMs and plays important roles in many disease processes, such as cancer. Identification of glycopeptides and glycans is essential to better understand the functions and bioactivities of glycoproteins. The progress of this study is mainly hindered by the lack of algorithms for intact glycopeptide characterization. We propose GlycoMaster DB in Chapter 6 to fulfil this urgent requirement on *N*-linked glycopeptides. GlycoMaster DB can analyze on a large-scale MS/MS dataset obtained from a biological sample with glycoproteins being either enriched or not, and from either HCD/ETD or HCD-only fragmentation. It enables the simultaneous identification of glycopeptide sequences and *N*-linked glycan composition from a user-specified protein database and a pre-configured *N*-linked

glycan database, respectively. Testing on four datasets demonstrates the promising performance of the software.

Protein quantification provides the information of protein quantity changes and assists in discoveries of important biomarkers in disease studies. Matching the peptide features extracted from different datasets is a crucial step to calculate the protein abundance ratios. Heuristic approaches have been proposed in previous research but none of them has yet claimed a clear optimization goal. In Chapter 7, a combinatorial problem, maximum peptide feature matching, is formulated and proven to be *NP*-hard. Practical algorithms are presented to solve the problem approximately in polynomial time and can help determine an upper-bound and a lower-bound of the optimal solution. The performances of our algorithms also compare favorably to other existing methods.

8.2 Future Work

Our future work will focus on providing algorithmic solutions for computational challenges encountered in mass spectrometry-based proteomics, especially glycoproteomics.

In the dissertation we introduce two approaches to improve the performance of *de novo* sequencing, and a new approach can be proposed for better identification of PTMs. When the number of PTM types increases, the accuracy of DeNovoPTM's result does not degrade too much in contrast to the conventional algorithms, but it is still not confident enough for practical use. More information is required to precisely identify the PTMs, which inspires the possible application of the DeNovoPTM algorithm on spectral data obtained from multiple fragmentation methods. Different fragmentation patterns can confirm each other on the discrimination between true and spurious fragmentation sites. It can be foreseeable that this combination can increase the confidence of PTMs identified from *de novo* sequencing.

Glycoprotein characterization is an urgent task in the emerging computational glycoproteomics. Several open problems need to be addressed in the near future. The first challenge is to refine the scoring scheme for the combination of spectra obtained from two types of fragmentation methods. It has been proven that two commonly used fragmentation methods (HCD and ETD) can fragment a glycopeptide with significantly different mechanisms. Therefore, utilization of both types of spectra can

simultaneously identify glycopeptides and glycan composition with higher confidence. Our preliminary experiments have shown that a straightforward scoring scheme can improve the identification from HCD/ETD spectral data. However, a sophisticated scoring model considering multiple types of spectra, as well as an improved experimental strategy, is strongly required for more confident characterization.

The second challenge is to construct and maintain a spectrum library of glycopeptides for MS/MS data analysis. This concept is similar to spectral library search: searching a given spectrum in a spectral library constituted by previously identified spectra, and using the best matched one to interpret the given spectrum. Such a spectral library is currently not available because of the lack of automated glycopeptide identification tools, and our GlycoMaster DB can fulfill the requirement. This research area is in its infancy and the library has the significant potential to accelerate the progress of the emerging glycoproteomics research.

The third challenge is to interpret the spectral data generated by *O*-linked glycoproteins. MS/MS spectra produced by *O*-linked glycopeptides are different from the ones by *N*-linked glycopeptides, and thus algorithms designed for *N*-linked glycopeptide characterization need to be cautiously revised. A large-scale analysis of glycoproteome in a biological sample demands a universal framework that characterizes both *N*-linked and *O*-linked glycopeptides, instead of using separated approaches. Furthermore, an effective statistical model is also required to control false positives.

In addition, the determination of glycan structures, rather than glycan composition, is another non-trivial extension of the glycopeptide characterization problem and demands experimental and computational solutions.

References

- [1] ABRF Delta Mass database. <http://www.abrf.org/index.cfm/dm.home>, 2009.
- [2] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.
- [3] William R. Alley, Yehia Mechref, and Milos V. Novotny. Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. *Rapid Communications in Mass Spectrometry*, 23(1):161–170, 2009.
- [4] N. Leigh Anderson and Norman G. Anderson. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19(11):1853–1861, 1998.
- [5] Kiyoko F. Aoki-Kinoshita. An introduction to bioinformatics for glycomics research. *PLoS Computational Biology*, 4(5):e1000075, 2008.
- [6] Rolf Apweiler, Henning Hermjakob, and Nathan Sharon. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta (BBA)–General Subjects*, 1473(1):4–8, 1999.
- [7] Manor Askenazi, Nuno. Bandeira, Robert J. Chalkley, Karl R. Clauser, Eric W. Deutsch, Henry H. N. Lam, W. Hayes McDonald, Thomas A. Neubert, Paul A. Rudnick, and Lennart Martens. iPRG 2011: a study on the identification of electron transfer dissociation (ETD) mass spectra. *Journal of Biomolecular Techniques*, 22(Supplement):S20, 2011.

REFERENCES

- [8] Jacques U. Baenziger. A major step on the road to understanding a unique posttranslational modification and its role in a genetic disease. *Cell*, 113(4):421–422, 2003.
- [9] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O’Donovan, Nicole Redaschi, and Lai-Su L. Yeh. The universal protein resource (UniProt). *Nucleic Acids Research*, 35(Database issue):D193–D197, 2007.
- [10] Nuno Bandeira, Dekel Tsur, Ari Frank, and Pavel A. Pevzner. Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences*, 104(15):6140–6145, 2007.
- [11] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical & Bioanalytical Chemistry*, 389(4):1017–1031, 2007.
- [12] Christopher Barner-Kowollik, Till Gruending, Jana Falkenhagen, and Steffen Weidner. *Mass spectrometry in polymer chemistry*. John Wiley & Sons, 2011.
- [13] Matthew Bellew, Marc Coram, Matthew Fitzgibbon, Mark Igra, Tim Randolph, Pei Wang, Damon May, Jimmy Eng, Ruihua Fang, Chenwei Lin, Jinzhi Chen, Jeffrey Goodlett, David amd Whiteaker, Amanda Paulovich, and Martin McIntosh. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, 22(15):1902–1909, 2006.
- [14] Marshall Bern, Yuhan Cai, and David Goldberg. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Analytical Chemistry*, 79(4):1393–1400, 2007.
- [15] Marshall Bern, Brett S. Phinney, and David Goldberg. Reanalysis of *Tyrannosaurus rex* mass spectra. *Journal of Proteome Research*, 8(9):4328–4332, 2009.
- [16] Andreas Bertsch, Andreas Leinenbach, Anton Pervukhin, Markus Lubeck, Ralf Hartmer, Carsten Baessmann, Yasser Abbas Elnakady, Rolf Müller, Sebastian

- Böcker, Christian G. Huber, and Oliver Kohlbacher. De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis*, 30(21):3736–3747, 2009.
- [17] Swapnil Bhatia, Yong Kil, Beatrix Ueberheide, Brian Chait, Lemmuel Tayo, Lourdes Cruz, Bingwen Lu, John R. Yates III, and Marshall Bern. Constrained de novo sequencing of peptides with application to conotoxins. In *Proceedings of the 15th Research in Computational Molecular Biology*, pages 16–30. Springer, 2011.
- [18] Walter P. Blackstock and Malcolm P. Weir. Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*, 17(3):121, 1999.
- [19] Patricie Burda and Markus Aebi. The dolichol pathway of n-linked glycosylation. *Biochimica et Biophysica Acta (BBA)–General Subjects*, 1426(2):239–257, 1999.
- [20] Dan Bylund, Rolf Danielsson, Gunnar Malmquist, and Karin E. Markides. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *Journal of Chromatography A*, 961(2):237–244, 2002.
- [21] Salvatore Cappadona, Peter R. Baker, Pedro R. Cutillas, Albert J. R. Heck, and Bas van Breukelen. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino Acids*, pages 1–22, 2012.
- [22] Alessio Ceroni, Kai Maass, Hildegard Geyer, Rudolf Geyer, Anne Dell, and Stuart M. Haslam. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *Journal of Proteome Research*, 7(4):1650–1659, 2008.
- [23] Robert J. Chalkley. When target-decoy false discovery rate estimations are inaccurate and how to spot instances. *Journal of Proteome Research*, 12(2):1062–1064, 2013.
- [24] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

REFERENCES

- [25] R. Graham Cooks. Collision-induced dissociation: Readings and commentary. *Journal of Mass Spectrometry*, 30(9):1215–1221, 1995.
- [26] Catherine A. Cooper, Elisabeth Gasteiger, and Nicolle H. Packer. GlycoMod – a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics*, 1(2):340–349, 2001.
- [27] Robertson Craig and Ronald C. Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry*, 17(20):2310–2316, 2003.
- [28] Robertson Craig and Ronald C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.
- [29] Robertson Craig, John P. Cortens, David Fenyo, and Ronald C. Beavis. Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research*, 5(8):1843–1849, 2006.
- [30] David M. Creasy and John S. Cottrell. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics*, 2(10):1426–1434, 2002.
- [31] David M. Creasy and John S. Cottrell. Unimod: protein modifications for mass spectrometry. *Proteomics*, 4(6):1534–1536, 2004.
- [32] Richard D. Cummings. The repertoire of glycan determinants in the human glycome. *Molecular BioSystems*, 5(10):1087–1104, 2009.
- [33] David C. Dallas, William F. Martin, Serenus Hua, and J. Bruce German. Automated glycopeptide analysis – review of current state and future directions. *Briefings in Bioinformatics*, 2012.
- [34] Vlado Dančik, Theresa A. Addona, Karl R. Clauser, James E. Vath, and Pavel A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.
- [35] Ritendra Datta and Marshall Bern. Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. *Journal of Computational Biology*, 16(8):1169–1182, 2009.
- [36] Bruno Domon and Ruedi Aebersold. Mass spectrometry and protein analysis. *Science*, 312(5771):212–217, 2006.

-
- [37] Bruno Domon and Catherine E. Costello. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal*, 5(4):397–409, 1988.
- [38] Scott Doubet, Klaus Bock, Dana Smith, Alan Darvill, and Peter Albersheim. The complex carbohydrate structure database. *Trends in Biochemical Sciences*, 14(12):475–477, 1989.
- [39] Mark W. Duncan, Ruedi Aebersold, and Richard M. Caprioli. The pros and cons of peptide-centric proteomics. *Nature Biotechnology*, 28(7):659–664, 2010.
- [40] Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007.
- [41] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.
- [42] Martin Ethier, Julian A. Saba, Maureen Spearman, Oleg Krokhin, Michael Butler, Werner Ens, Kenneth G. Standing, and Helene Perreault. Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(24):2713–2720, 2003.
- [43] Attila Felinger. *Data analysis and signal processing in chromatography*, volume 51. Elsevier, Amsterdam, 1998.
- [44] Bernd Fischer, Jonas Grossmann, Volker Roth, Wilhelm Gruissem, Sacha Baginsky, and Joachim M. Buhmann. Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics*, 22(14):e132–e140, 2006.
- [45] Ari Frank and Pavel A. Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77(4):964–973, 2005.
- [46] Michael L. Fredman and Robert E. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM (JACM)*, 34(3):596–615, 1987.

REFERENCES

- [47] Hudson H. Freeze and Markus Aebi. Altered glycan structures: the molecular basis of congenital disorders of glycosylation. *Current Opinion in Structural Biology*, 15(5):490–498, 2005.
- [48] Barbara E Frewen, Gennifer E Merrihew, Christine C Wu, William Stafford Noble, and Michael J MacCoss. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical Chemistry*, 78(16):5678–5684, 2006.
- [49] Michael R. Garey and David S. Johnson. *Computers and intractability: a guide to the theory of NP-completeness*. WH Freeman and Company, New York, 1979.
- [50] Sara P. Gaucher, Jeff Morrow, and Julie A. Leary. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Analytical Chemistry*, 72(11):2331–2336, 2000.
- [51] Ylva Gavel and Gunnar von Heijne. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Engineering*, 3(5):433–442, 1990.
- [52] Lewis Y. Geer, Sanford P. Markey, Jeffrey A. Kowalak, Lukas Wagner, Ming Xu, Dawn M. Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H. Bryant. Open mass spectrometry search algorithm. *Journal of Proteome Research*, 3(5):958–964, 2004.
- [53] David Goldberg, Marshall Bern, Simon Parry, Mark Sutton-Smith, Maria Panico, Howard R. Morris, and Anne Dell. Automated N-glycopeptide identification using a combination of single- and tandem-MS. *Journal of Proteome Research*, 6(10):3995–4005, 2007.
- [54] David Goldberg, Mark Sutton-Smith, James Paulson, and Anne Dell. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics*, 5(4):865–875, 2005.
- [55] Donald J. Graves, Bruce L. Martin, and Jerry H. Wang. *Co- and post-translational modification of proteins: chemical principles and biological effects*, volume 24. Oxford University Press, New York, 1994.

-
- [56] Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A. Pevzner. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry*, 22(7):1111–1120, 2011.
- [57] Steven P. Gygi, Beate Rist, Scott A. Gerber, Frantisek Turecek, Michael H. Gelb, and Ruedi Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17(10):994–999, 1999.
- [58] Per Hägglund, Jakob Bunkenborg, Felix Elortza, Ole Nørregaard Jensen, and Peter Roepstorff. A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation. *Journal of Proteome Research*, 3(3):556–566, 2004.
- [59] Per Hägglund, Rune Matthiesen, Felix Elortza, Peter Højrup, Peter Roepstorff, Ole Nørregaard Jensen, and Jakob Bunkenborg. An enzymatic deglycosylation scheme enabling identification of core fucosylated N-glycans and O-glycosylation site mapping of human plasma proteins. *Journal of Proteome Research*, 6(8):3021–3031, 2007.
- [60] Xi Han, Lin He, Lei Xin, Baozhen Shan, and Bin Ma. PeaksPTM: mass spectrometry-based identification of peptides with unspecified modifications. *Journal of Proteome Research*, 10(7):2930–2936, 2011.
- [61] Yonghua Han, Bin Ma, and Kaizhong Zhang. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *Journal of Bioinformatics and Computational Biology*, 3(03):697–716, 2005.
- [62] Kosuke Hashimoto and Minoru Kanehisa. KEGG GLYCAN for integrated analysis of pathways, genes, and structures. *Experimental Glycoscience*, pages 441–444, 2008.
- [63] Lin He, Xi Han, and Bin Ma. De novo sequencing with limited number of post-translational modifications per peptide. *Journal of Bioinformatics and Computational Biology*, 2013.
- [64] Lin He and Bin Ma. Adepts: advanced peptide de novo sequencing with a pair of tandem mass spectra. *Journal of Bioinformatics and Computational Biology*, 8(06):981–994, 2010.

REFERENCES

- [65] Albert J. R. Heck and Jeroen Krijgsveld. Mass spectrometry-based quantitative proteomics. *Expert Review of Proteomics*, 1(3):317–326, 2004.
- [66] Franz Hillenkamp, Michael Karas, Ronald C. Beavis, and Brian T. Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, 63(24):1193A–1203A, 1991.
- [67] Wade M. Hines, Arnold M. Falick, Alma L. Burlingame, and Bradford W. Gibson. Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *Journal of the American Society for Mass Spectrometry*, 3(4):326–336, 1992.
- [68] Edmond Hoffmann and Vincent Stroobant. Mass spectrometry: principles and applications, 2007.
- [69] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>, 2003.
- [70] Barbara Imperiali and Sarah E. O’Connor. Effect of N-linked glycosylation on glycopeptide and glycoprotein structure. *Current Opinion in Chemical Biology*, 3(6):643, 1999.
- [71] Navdeep Jaitly, Matthew E. Monroe, Vladislav A. Petyuk, Therese R. W. Clauss, Joshua N. Adkins, and Richard D. Smith. Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Analytical Chemistry*, 78(21):7397–7409, 2006.
- [72] Philip Jones, Richard G. Côté, Sang Yun Cho, Sebastian Klie, Lennart Martens, Antony F. Quinn, David Thorneycroft, and Henning Hermjakob. PRIDE: new developments and new datasets. *Nucleic Acids Research*, 36(suppl 1):D878–D883, 2008.
- [73] Philip Jones, Richard G. Côté, Lennart Martens, Antony F. Quinn, Chris F. Taylor, William Derache, Henning Hermjakob, and Rolf Apweiler. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Research*, 34(suppl 1):D659–D663, 2006.

-
- [74] Hiren J. Joshi, Mathew J. Harrison, Benjamin L. Schulz, Catherine A. Cooper, Nicole H. Packer, and Niclas G. Karlsson. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics*, 4(6):1650–1664, 2004.
- [75] Lukas Käll, John D. Storey, Michael J. MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research*, 7(01):29–34, 2008.
- [76] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl 1):D277–D280, 2004.
- [77] Andrew Keller, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74(20):5383–5392, 2002.
- [78] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J. Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert J. R. Heck, and Pavel A. Pevzner. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics*, 9(12):2840–2852, 2010.
- [79] Sangtae Kim, Seungjin Na, Ji Woong Sim, Heejin Park, Jaeho Jeong, Hokeun Kim, Younghwan Seo, Jawon Seo, Kong-Joo Lee, and Eunok Paek. MODⁱ: a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra. *Nucleic Acids Research*, 34(suppl 2):W258–W263, 2006.
- [80] Young J. Kim and Ajit Varki. Perspectives on the significance of altered glycosylation of glycoproteins in cancer. *Glycoconjugate Journal*, 14(5):569–576, 1997.
- [81] Marc Kirchner, Benjamin Saussen, Hanno Steen, Judith A. J. Steen, and Fred A. Hamprrecht. Amsrpm: robust point matching for retention time alignment of LC/MS data with R. *Journal of Statistical Software*, 18(4), 2007.

REFERENCES

- [82] John Klimek, James S. Eddes, Laura Hohmann, Jennifer Jackson, Amelia Peterson, Simon Letarte, Philip R. Gafken, Jonathan E. Katz, Parag Mallick, Hookeun Lee, Alexander Schmidt, Reto Ossola, Jimmy K. Eng, Ruedi Aebersold, and Daniel B. Martin. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *Journal of Proteome Research*, 7(01):96–103, 2007.
- [83] Henry Lam, Eric W. Deutsch, James S. Eddes, Jimmy K. Eng, Nichole King, Stephen E. Stein, and Ruedi Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, 7(5):655–667, 2007.
- [84] Brian L. LaMarche, Kevin L. Crowell, Navdeep Jaitly, Vladislav A. Petyuk, Anuj R. Shah, Ashoka D. Polpitiya, John D. Sandoval, Gary R. Kiebel, Matthew E. Monroe, Stephen J. Callister, Thomas O. Metz, Gordon A. Anderson, and Richard D. Smith. MultiAlign: a multiple LC-MS analysis tool for targeted omics analysis. *BMC Bioinformatics*, 14(1):49, 2013.
- [85] Eva Lange, Clemens Gröpl, Ole Schulz-Trieglaff, Andreas Leinenbach, Christian Huber, and Knut Reinert. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics*, 23(13):i273–i281, 2007.
- [86] Eva Lange, Ralf Tautenhahn, Steffen Neumann, and Clemens Gröpl. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, 9(1):375, 2008.
- [87] Anthony J. Lapadula, Philip J. Hatcher, Andy J. Hanneman, David J. Ashline, Hailong Zhang, and Vernon N. Reinhold. Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MS^n data. *Analytical Chemistry*, 77(19):6271–6279, 2005.
- [88] Xiao-jun Li, C. Yi Eugene, Christopher J. Kemp, Hui Zhang, and Ruedi Aebersold. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Molecular & Cellular Proteomics*, 4(9):1328–1340, 2005.
- [89] Hao Lin, Lin He, and Bin Ma. A combinatorial approach to the peptide feature matching problem for label-free quantification. *Bioinformatics*, 2013.

-
- [90] Xiaowen Liu, Yonghua Han, Denis Yuen, and Bin Ma. Automated protein (re) sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics*, 25(17):2174–2180, 2009.
- [91] Xiaowen Liu, Baozhen Shan, Lei Xin, and Bin Ma. Better score function for peptide identification with ETD MS/MS spectra. *BMC Bioinformatics*, 11(Suppl 1):S4, 2010.
- [92] Klaus K. Lohmann and Claus-Wilhelm von der Lieth. GLYCO-FRAGMENT: a web tool to support the interpretation of mass spectra of complex carbohydrates. *Proteomics*, 3(10):2028–2035, 2003.
- [93] Klaus K. Lohmann and Claus-Wilhelm von der Lieth. GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Research*, 32(suppl 2):W261–W266, 2004.
- [94] Bingwen Lu and Ting Chen. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 10(1):1–12, 2003.
- [95] Thomas Lütteke, Andreas Bohne-Lang, Alexander Loss, Thomas Goetz, Martin Frank, and Claus-Wilhelm von der Lieth. GLYCOSCIENCES.de: an internet portal to support glycomics and glycobiology research. *Glycobiology*, 16(5):71R–81R, 2006.
- [96] Bin Ma. Challenges in computational analysis of mass spectrometry data for proteomics. *Journal of Computer Science & Technology*, 25(1):107–123, 2010.
- [97] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–2342, 2003.
- [98] Bin Ma, Kaizhong Zhang, and Chengzhi Liang. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *Journal of Computer and System Sciences*, 70(3):418–430, 2005.
- [99] Kai Maass, René Ranzinger, Hildegard Geyer, Claus-Wilhelm von der Lieth, and Rudolf Geyer. “Glyco-peakfinder” – de novo composition analysis of glycoconjugates. *Proteomics*, 7(24):4435–4444, 2007.

REFERENCES

- [100] Matthias Mann and Ole N. Jensen. Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21(3):255–261, 2003.
- [101] Matthias Mann and Matthias Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66(24):4390–4399, 1994.
- [102] Matthias Mann and Matthias Wilm. Electrospray mass spectrometry for protein characterization. *Trends in Biochemical Sciences*, 20(6):219–224, 1995.
- [103] Arivusudar Marimuthu, Robert N. O’Meally, Raghothama Chaerkady, Yashwanth Subbannayya, Vishalakshi Nanjappa, Praveen Kumar, Dhanashree S. Kelkar, Sneha M. Pinto, Rakesh Sharma, Santosh Renuse, Renu Goel, Rita Christopher, Bernard Delanghe, Robert N. Cole, H. C. Harsha, and Akhilesh Pandey. A comprehensive map of the human urinary proteome. *Journal of Proteome Research*, 10(6):2734–2743, 2011.
- [104] Anoop M. Mayampurath, Yin Wu, Zaneer M. Segu, Yehia Mechref, and Haixu Tang. Improving confidence in detection and characterization of protein N-glycosylation sites and microheterogeneity. *Rapid Communications in Mass Spectrometry*, 25(14):2007–2019, 2011.
- [105] W. Hayes McDonald and John R. Yates III. Shotgun proteomics: integrating technologies to answer biological questions. *Current Opinion in Molecular Therapeutics*, 5(3):302, 2003.
- [106] Leann M. Mikesch, Beatrix Ueberheide, An Chi, Joshua J. Coon, John E. P. Syka, Jeffrey Shabanowitz, and Donald F. Hunt. The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta*, 1764(12):1811–1822, 2006.
- [107] Lijuan Mo, Debojyoti Dutta, Yunhu Wan, and Ting Chen. MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Analytical Chemistry*, 79(13):4870–4878, 2007.
- [108] Marcin Mucha and Piotr Sankowski. Maximum matchings via Gaussian elimination. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 248–255. IEEE, 2004.

-
- [109] Lukas N. Mueller, Oliver Rinner, Alexander Schmidt, Simon Letarte, Bernd Bodenmiller, Mi-Youn Brusniak, Olga Vitek, Ruedi Aebersold, and Markus Müller. SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*, 7(19):3470–3480, 2007.
- [110] Nagarjuna Nagaraj, Nils Alexander Kulak, Juergen Cox, Nadin Neuhauser, Korbinian Mayr, Ole Hoerning, Ole Vorm, and Matthias Mann. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Molecular & Cellular Proteomics*, 11(3), 2012.
- [111] Alexey I. Nesvizhskii, Olga Vitek, and Ruedi Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods*, 4(10):787–797, 2007.
- [112] Kazuaki Ohtsubo and Jamey D. Marth. Glycosylation in cellular mechanisms of health and disease. *Cell*, 126(5):855–867, 2006.
- [113] William M. Old, Karen Meyer-Arendt, Lauren Aveline-Wolf, Kevin G. Pierce, Alex Mendoza, Joel R. Sevinisky, Katheryn A. Resing, and Natalie G. Ahn. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Molecular & Cellular Proteomics*, 4(10):1487–1502, 2005.
- [114] Jesper V. Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712, 2007.
- [115] Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics*, 1(5):376–386, 2002.
- [116] Sheng Pan, Ruedi Aebersold, Ru Chen, John Rush, David R. Goodlett, Martin W. McIntosh, Jing Zhang, and Teresa A. Brentnall. Mass spectrometry based targeted protein quantification: methods and applications. *Journal of Proteome Research*, 8(2):787–797, 2008.

REFERENCES

- [117] Sheng Pan, Ru Chen, Ruedi Aebersold, and Teresa A. Brentnall. Mass spectrometry based glycoproteomics from a proteomics perspective. *Molecular & Cellular Proteomics*, 10(1), 2011.
- [118] David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.
- [119] Emanuel F. Petricoin, Kathryn C. Zoon, Elise C. Kohn, J. Carl Barrett, and Lance A. Liotta. Clinical proteomics: translating benchside promise into bedside reality. *Nature Reviews Drug Discovery*, 1(9):683–695, 2002.
- [120] Sergey Pevtsov, Irina Fedulova, Hamid Mirzaei, Charles Buck, and Xiang Zhang. Performance evaluation of existing de novo sequencing algorithms. *Journal of Proteome Research*, 5(11):3018–3028, 2006.
- [121] Tomáš Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Orešič. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(1):395, 2010.
- [122] Katharina Podwojski, Arno Fritsch, Daniel C. Chamrad, Wolfgang Paul, Barbara Sitek, Kai Stühler, Petra Mutzel, Christian Stephan, Helmut E. Meyer, Wolfgang Urfer, Katja Ickstadt, and Jörg Rahnenfhrer. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics*, 25(6):758–764, 2009.
- [123] Dragan Radulovic, Salomeh Jelveh, Soyoung Ryu, T. Guy Hamilton, Eric Foss, Yongyi Mao, and Andrew Emili. Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Molecular & cellular proteomics*, 3(10):984–997, 2004.
- [124] Rahul Raman, Maha Venkataraman, Subu Ramakrishnan, Wei Lang, S. Ragu-ram, and Ram Sasisekharan. Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, 16(5):82R–90R, 2006.
- [125] René Ranzinger, Stephan Herget, Thomas Wetter, and Claus-Wilhelm Von Der Lieth. GlycomeDB – integration of open-access carbohydrate structure databases. *BMC Bioinformatics*, 9(1):384, 2008.

-
- [126] René Ranzinger, Kai Maaß, and Thomas Lütteke. Bioinformatics databases and applications available for glycobiology and glycomics. In *Functional and Structural Proteomics of Glycoproteins*, pages 59–90. Springer, 2011.
- [127] Franz F. Roos, Riko Jacob, Jonas Grossmann, Bernd Fischer, Joachim M Buhmann, Wilhelm Gruissem, Sacha Baginsky, and Peter Widmayer. PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. *Bioinformatics*, 23(22):3016–3023, 2007.
- [128] Julian Saba, Sucharita Dutta, Eric Hemenway, and Rosa Viner. Increasing the productivity of glycopeptides analysis by using higher-energy collision dissociation-accurate mass-product-dependent electron transfer dissociation. *International Journal of Proteomics*, 2012, 2012.
- [129] Mikhail M. Savitski, Michael L. Nielsen, Frank Kjeldsen, and Roman A. Zubarev. Proteomics-grade de novo sequencing approach. *Journal of Proteome Research*, 4(6):2348–2354, 2005.
- [130] Mikhail M. Savitski, Michael L. Nielsen, and Roman A. Zubarev. Modificomb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Molecular & Cellular Proteomics*, 5(5):935–948, 2006.
- [131] Scott and Peter Albersheim. Carbbank. *Glycobiology*, 2(6):505, 1992.
- [132] Nichollas E. Scott, Benjamin L. Parker, Angela M. Connolly, Jana Paulech, Alistair V. G. Edwards, Ben Crossett, Linda Falconer, Daniel Kolarich, Steven P. Djordjevic, Peter Højrup, Nicolle H. Packer, Martin R. Larsen, and Stuart J. Cordwell. Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of *Campylobacter jejuni*. *Molecular & Cellular Proteomics*, 10(2), 2011.
- [133] Brian C. Searle, Surendra Dasari, Mark Turner, Ashok P. Reddy, Dongseok Choi, Phillip A. Wilmarth, Ashley L. McCormack, Larry L. David, and Srinivasa R. Nagalla. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Analytical Chemistry*, 76(8):2220–2230, 2004.

REFERENCES

- [134] Baozhen Shan, Kaizhong Zhang, Bin Ma, Cunjie Zhang, and Gilles Lajoie. GlycoMaster – A software for interpretation of glycopeptides from MS/MS spectra. In *Proceedings of the 52nd ASMS Conference on Mass Spectrometry and Allied Topics*, 2004.
- [135] Ignat V. Shilov, Sean L. Seymour, Alpesh A. Patel, Alex Loboda, Wilfred H. Tang, Sean P. Keating, Christie L. Hunter, Lydia M. Nuwaysir, and Daniel 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. *Molecular & Cellular Proteomics*, 6(9):1638–1655, 2007.
- [136] Jeffrey C. Silva, Richard Denny, Craig A. Dorschel, Marc Gorenstein, Ignatius J. Kass, Guo-Zhong Li, Therese McKenna, Michael J. Nold, Keith Richardson, Phillip Young, and Scott Geromanos. Quantitative proteomic analysis by accurate mass retention time pairs. *Analytical Chemistry*, 77(7):2187–2200, 2005.
- [137] Charandeep Singh, Cleidiane G. Zampronio, Andrew J. Creese, and Helen J. Cooper. Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) Mass Spectrometry for the Analysis of N-Linked Glycoproteins. *Journal of Proteome Research*, 11(9):4517–4525, 2012.
- [138] Gordon W. Slysz, Erin S. Baker, Anuj R. Shah, Navdeep Jaitly, Gordon A. Anderson, and Richard D. Smith. The decontools framework: an application programming interface enabling flexibility in accurate mass and time tag workflows for proteomics and metabolomics. In *Proceedings of the 58th ASMS Conference on Mass Spectrometry and Allied Topics*, 2010.
- [139] Colin A Smith, J. Elizabeth, Grace O’Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006.
- [140] Sudhir Srivastava. Move over proteomics, here comes glycomics. *Journal of Proteome Research*, 7(5):1799–1799, 2008.
- [141] Marc Sturm, Andreas Bertsch, Clemens Gröpl, Andreas Hildebrandt, Rene Hussong, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, Alexandra Zerck, Knut Reinert, and Oliver Kohlbacher. OpenMS – an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9(1):163, 2008.

-
- [142] Danielle L. Swaney, Graeme C. McAlister, and Joshua J. Coon. Decision tree – driven tandem mass spectrometry for shotgun proteomics. *Nature Methods*, 5(11):959–964, 2008.
- [143] John E.P. Syka, Joshua J. Coon, Melanie J. Schroeder, Jeffrey Shabanowitz, and Donald F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences*, 101(26):9528–9533, 2004.
- [144] David L. Tabb, Anita Saraf, and John R. Yates III. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical Chemistry*, 75(23):6415–6421, 2003.
- [145] Haixu Tang, Yehia Mechref, and Milos V. Novotny. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics*, 21(suppl 1):i431–i439, 2005.
- [146] Stephen Tanner, Hongjun Shu, Ari Frank, Ling-Chi Wang, Ebrahim Zandi, Marc Mumby, Pavel A. Pevzner, and Vineet Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, 77(14):4626–4639, 2005.
- [147] Anthony L. Tarentino, Caroline M. Gomez, and Thomas H. Plummer Jr. Deglycosylation of asparagine-linked glycans by peptide: N-glycosidase F. *Biochemistry*, 24(17):4665–4671, 1985.
- [148] J. Alex Taylor and Richard S. Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, 73(11):2594–2604, 2001.
- [149] John F. Timms and Pedro R. Cutillas. Overview of quantitative LC-MS techniques for proteomics and activitomics. *Methods in Molecular Biology*, 658:19–45, 2010.
- [150] Chih-Chiang Tsou, Chia-Feng Tsai, Ying-Hao Tsui, Putty-Reddy Sudhir, Yi-Ting Wang, Yu-Ju Chen, Jeou-Yuan Chen, Ting-Yi Sung, and Wen-Lian Hsu. IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. *Molecular & Cellular Proteomics*, 9(1):131–144, 2010.

REFERENCES

- [151] Dekel Tsur, Stephen Tanner, Ebrahim Zandi, Vineet Bafna, and Pavel A. Pevzner. Identification of post-translational modifications by blind search of mass spectra. *Nature Biotechnology*, 23(12):1562–1567, 2005.
- [152] Mathias Vandenbergert, Sébastien Li-Thiao-Té, Hans-Michael Kaltenbach, Runxuan Zhang, Tero Aittokallio, and Benno Schwikowski. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics*, 8(4):650–672, 2008.
- [153] Ajit Varki. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, 3(2):97–130, 1993.
- [154] Ajit Varki, Richard D. Cummings, Jeffrey D. Esko, Hudson H. Freeze, Pamela Stanley, Carolyn R. Bertozzi, Gerald W. Hart, and Marilyn E. Etzler. *Essentials of Glycobiology. 2nd edition*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2009.
- [155] Claus-Wilhelm von der Lieth, Ana Ardá Freire, Dennis Blank, Matthew P. Campbell, Alessio Ceroni, David R Damerell, Anne Dell, Raymond A Dwek, Beat Ernst, Rasmus Fogh, et al. EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology*, 21(4):493–502, 2011.
- [156] Weixun Wang, Haihong Zhou, HUA Lin, Sushmita Roy, Thomas A. Shaler, Lander R. Hill, Scott Norton, Praveen Kumar, Markus Anderle, and Christopher H. Becker. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry*, 75(18):4818–4826, 2003.
- [157] Sebastian Wiese, Kai A. Reidegeld, Helmut E. Meyer, and Bettina Warscheid. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics*, 7(3):340–350, 2007.
- [158] Eric S. Witze, William M. Old, Katheryn A. Resing, and Natalie G. Ahn. Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, 4(10):798–806, 2007.
- [159] Finn Wold. In vivo chemical modification of proteins (post-translational modification). *Annual Review of Biochemistry*, 50(1):783–814, 1981.

-
- [160] Carrie L. Woodin, David Hua, Morgan Maxon, Kathryn R. Rebecchi, Eden P. Go, and Heather Desaire. GlycoPep Grader: a web-based utility for assigning the composition of N-linked glycopeptides. *Analytical Chemistry*, 84(11):4821–4829, 2012.
- [161] Mark R. Wormald and Raymond A. Dwek. Glycoproteins: glycan presentation and protein-fold stability. *Structure*, 7(7):R155–R160, 1999.
- [162] Manfred Wührer, M Isabel Catalina, André M Deelder, and Cornelis H. Hokke. Glycoproteomics based on tandem mass spectrometry of glycopeptides. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 849(1-2):115–128, 2007.
- [163] John R. Yates III. Database searching using mass spectrometry data. *Electrophoresis*, 19(6):893–900, 1998.
- [164] John R. Yates III, Jimmy K. Eng, Ashley L. McCormack, and David Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry*, 67(8):1426–1436, 1995.
- [165] Jianqiu Zhang, Elias Gonzalez, Travis Hestilow, William Haskins, and Yufei Huang. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Current Genomics*, 10(6):388, 2009.
- [166] Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A. Lajoie, and Bin Ma. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics*, 11(4), 2012.
- [167] Wei Zhang, Hong Wang, Lei Zhang, Jun Yao, and Pengyuan Yang. Large-scale assignment of N-glycosylation sites using complementary enzymatic deglycosylation. *Talanta*, 85(1):499–505, 2011.
- [168] Yaoyang Zhang, Bryan R Fonslow, Bing Shan, Moon-Chang Baek, and John R Yates III. Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews*, 113(4):2343–2394, 2013.
- [169] Wenhong Zhu, Jeffrey W. Smith, and Chun-Ming Huang. Mass spectrometry-based label-free quantitative proteomics. *BioMed Research International*, 2010, 2009.

REFERENCES

- [170] Zhikai Zhu, David Hua, Daniel F. Clark, Eden P. Go, and Heather Desaire. GlycoPep Detector: a tool for assigning mass spectrometry data of N-Linked glycopeptides on the basis of their electron transfer dissociation spectra. *Analytical Chemistry*, 85(10):5023–5032, 2013.
- [171] Roman A. Zubarev, David M. Horn, Einar K. Fridriksson, Neil L. Kelleher, Nathan A. Kruger, Mark A. Lewis, Barry K. Carpenter, and Fred W. McLafferty. Electron capture dissociation for structural characterization of multiply charged protein cations. *Analytical Chemistry*, 72(3):563–573, 2000.