

9-20-2016 12:00 AM

Algorithms for Glycan Structure Identification with Tandem Mass Spectrometry

Weiping Sun, *The University of Western Ontario*

Supervisor: Kaizhong Zhang, *The University of Western Ontario*

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

© Weiping Sun 2016

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>



Part of the [Bioinformatics Commons](#), and the [Theory and Algorithms Commons](#)

Recommended Citation

Sun, Weiping, "Algorithms for Glycan Structure Identification with Tandem Mass Spectrometry" (2016).
Electronic Thesis and Dissertation Repository. 4105.
<https://ir.lib.uwo.ca/etd/4105>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Abstract

Glycosylation is a frequently observed post-translational modification (PTM) of proteins. It has been estimated over half of eukaryotic proteins in nature are glycoproteins. Glycoprotein analysis plays a vital role in drug preparation. Thus, characterization of glycans that are linked to proteins has become necessary in glycoproteomics. Mass spectrometry has become an effective analytical technique for glycoproteomics analysis because of its high throughput and sensitivity. The large amount of spectral data collected in a mass spectrometry experiment makes manual interpretation impossible and requires effective computational approaches for automated analysis. Different algorithmic solutions have been proposed to address the challenges in glycoproteomics analysis based on mass spectrometry. However, new algorithms that can identify intact glycopeptides are still demanded to improve result accuracy.

In this research, a glycan is represented as a rooted unordered labelled tree and we focus on developing effective algorithms to determine glycan structures from tandem mass spectra. Interpreting the tandem mass spectra of glycopeptides with a *de novo* sequencing method is essential to identifying novel glycan structures. Thus, we mathematically formulated the glycan *de novo* sequencing problem and propose a heuristic algorithm for glycan *de novo* sequencing from HCD tandem mass spectra of glycopeptides.

Characterizing glycans from MS/MS with a *de novo* sequencing method requires high-quality mass spectra for accurate results. The database search method usually has the ability to obtain more reliable results since it has the assistance of glycan structural information. Thus, we propose a *de novo* sequencing assisted database search method, GlycoNovoDB, for mass spectra interpretation.

Keywords: Tandem mass spectrometry, glycan identification, glycopeptide, glycosylation

Acknowledgements

It has been four years since I came to Western pursuing this doctorate. I would like to express my sincerest appreciation to my supervisor, Dr. Kaizhong Zhang. His intelligence, inspiration, consideration, and patience makes it a great pleasure to study under his instruction over these years. I deeply realized that I can hardly complete my doctoral study without his unconditional supports and warm encourage.

I would like to send my gratitude to the supervisory committees, Dr. Lucian Ilie and Dr. Mark Daley, for spending their precious time on reading my Topic/Survey Proposal and providing valuable suggestions for my initial work. Also, I would like to thank my thesis examiners, (in alphabetical order) Dr. Lucian Ilie, Dr. Lila Kari, Dr. Shawn Li and Dr. Fangxiang Wu for their helpful advices concerning this work.

I would also like to thank all my dear colleagues, Dr. Weiming Li, Dr. Lin He, Dr. Yi Liu, Dr. Yan Yan, Zhewei Liang, Fang Han, Yu Shan, Qin Dong, Yiwei Li, Mike Molnar, and Nilesh Khiste, for their company during the past four years. They make my life full of gladness and help me relieve stress and homesick, which make us to be lifetime friends.

Many thanks to all of our collaborators, Dr. Bin Ma and Dr. Gilles A. Lajoie, for their insightful suggestions and positive assistance.

Specially, I am extremely thankful to my lovely family for their unconditional love and continuous encouragement. And, I would like to express my thanks to my dear partner, Yi. No matter what happened he always accompany me by my side, which makes me full of courage and passion to conquer any difficulties. Last but not least, I would like to thank my pet cat, Xiaohua, for her loveliness and trust that make my life lively and colourful.

*To my dear parents and partner, for their always being supportive during my
doctoral adventure*

Contents

Abstract	ii
Acknowledgements	iii
List of Figures	vii
List of Tables	ix
List of Abbreviations and Terminologies	x
List of Algorithms	xi
1 Introduction	1
2 Background	5
2.1 Fundamentals of Glycoproteins	5
2.1.1 Carbohydrates	5
2.1.2 Amino Acids	9
2.1.3 Glycans and Glycoproteins	10
2.2 Mass Spectrometry Technology	11
2.2.1 Mass Spectrometers and Configuration	12
2.2.2 Tandem Mass Spectrometry	15
2.3 Mass Spectrometry Based Glycoproteomics Analysis	17
2.3.1 General Strategies	17
2.3.2 Fragmentation Patterns	19
2.4 Computational Approaches for Interpreting MS/MS Spectra of Glycopeptides .	21
2.4.1 Database Search	22
2.4.2 Glycan <i>De Novo</i> Sequencing	24
3 Glycan <i>De Novo</i> Sequencing from HCD Spectra	28
3.1 Notations and Problem Definition	30
3.1.1 Basic Notation	30
3.1.2 Mass Representation of Ion Fragments	31
3.1.3 Problem Definition	33
3.1.4 Scoring Function	34
3.2 Methods	35
3.2.1 Peptide Mass Inference	35

3.2.2	Algorithm for Candidate Generation	37
3.2.3	Algorithm for Elimination of Duplicate Trees	43
3.2.4	Complexity Analysis	46
3.2.5	Re-evaluation Scheme	48
3.3	Experiments and Results	49
3.3.1	Datasets	49
3.3.2	Experimental Results	50
3.4	An Improved Result	53
3.5	Conclusion and Discussion	54
4	Glycan Structure Identification using Database Search and <i>De Novo</i> Sequencing	56
4.1	Preliminaries	57
4.1.1	Notations and Problem Formulation	57
4.1.2	Scoring Function	59
4.2	Methods	60
4.2.1	Glycan Candidates Selection and Raw Score Calculation	61
4.2.2	Algorithm for Filtration by <i>De Novo</i> Sequencing	62
4.2.3	Algorithm for Calculating Similarity Between Labelled Unordered Trees	64
4.2.4	Complexity Analysis	68
4.3	Experiments and Results	69
4.3.1	Datasets	69
4.3.2	Experimental Results	70
4.4	Recent Improvements	72
4.4.1	Methods	73
4.4.2	Experimental Results	75
4.5	Conclusion	78
5	Conclusion and Future Work	80
5.1	Conclusion	80
5.2	Future Work	82
	Bibliography	85
	Curriculum Vitae	95

List of Figures

2.1	The linear form and ring form of a monosaccharide (mannose).	6
2.2	Two glucose molecules combine to form a maltose and release a water. The hemiacetal carbon atom (C_1) is called anomeric carbon.	8
2.3	An example of tetrasaccharide and its tree representation.	8
2.4	N-linked and O-linked glycoproteins.	11
2.5	Methods of calculating mass resolving power: (a) calculation via 5% and 10% valley definition. (b) calculation via full width at half maximum (FWHM) definition.	13
2.6	An example of a visualized mass spectrum. The x-axis of the mass spectrum represents mass-to-charge ratio (m/z), and y-axis represents signal intensity of the ions.	15
2.7	Schematic of tandem mass spectrometry. The sample is ionized in the mass ionizer first, and then analyzed by the first mass analyzer. Ions of interest (<i>i.e.</i> precursor ion) are selected from the generated survey scan and fragmented into product ions, which are analyzed by the second mass analyzer. Finally, MS/MS spectrum is produced.	16
2.8	The strategies of mass spectrometry based glycoproteomics analysis. One strategy is to isolate glycans from glycopeptides and then generate mass spectra of the released glycans and deglycosylated peptides separately. The other strategy produces mass spectra directly from intact glycopeptides.	18
2.9	Examples of fragmented ions of a peptide. This peptide consists of four amino acids. Six basic types of fragment ions, including <i>a</i> -, <i>b</i> -, <i>c</i> -, <i>x</i> -, <i>y</i> -, and <i>z</i> -ions, generated by cleavages on the peptide backbone.	20
2.10	Examples of fragmented ions of a glycan. This glycan contains three monosaccharides. <i>A</i> -, <i>B</i> -, <i>C</i> -ions and <i>X</i> -, <i>Y</i> -, <i>Z</i> -ions are shown in the figure, where <i>A/X</i> -ions generated from cross ring cleavages. It is worth noticing that not all the ions are marked in the figure.	21
3.1	A glycopeptide and its tree representation. (a) A glycopeptide. The glycan attached to the peptide consists of four monosaccharides. (b) The abstract tree representation of the glycopeptide. Each monosaccharide is abstracted as a node in the tree representation. Each glycosidic bond between two monosaccharide is represented as a linkage of the tree.	32
3.2	An example for core structure in an N-linked glycopeptide. The two GlcNAc linked to a branched Man triad in the dotted rectangle is the core structure. . . .	36

3.3	An example for new possible trees generated by adding a node g to a glycan tree T . The added node is shaded in grey.	39
3.4	An example demonstrating the sets of $RPST(T)$ and $RPST(T, v)$ for the given glycan tree structure T . The node v in the tree T is shaded grey. There are five different root-preserving subtrees of T , in which three of them contain the node v and include a path from root to the node v	41
3.5	(a) Strings assigned by isomorphic tree elimination algorithm. The final string associated with the tree is $2(2(1(2(1(3))))(1(2(1(4))))))$. (b) Integers assigned to different types of monosaccharides.	45
3.6	Comparison of the top two glycan structures reported by GlycoMaster DB and our method for a certain spectrum. (a) Ranked No.1 and No.32 in the results of GlycoMaster DB and our method respectively. (b) Top ranked result in our method. (c) Ranked second place in both methods.	51
3.7	Comparison of two pairs of glycan structures identified by GlycoMaster DB and our method respectively. (a1) and (a2) are the top ranked glycans identified from the same spectrum and (b1) and (b2) are identified from another spectrum. Both (a1) and (b1) are the results reported by GlycoMaster DB, while (a2) and (b2) are reported by our method.	52
4.1	An example of permutation between $F_1[i]$ and $F_2[j]$, where the two subtrees $T_2[j_1]$ and $T_2[j_2]$ in $F_2[j]$ are mapped to the two subtrees in $F_1[i]$ which can reach the minimum cost, say $T_1[i_2]$ and $T_1[i_4]$ respectively, and all the other subtrees in $F_1[i]$ are deleted.	68
4.2	An example of glycans identified from the same spectrum that share the same top score in the results reported by GlycoMaster DB. However, in the results of GlycoNovoDB, glycan (a) is ranked higher than glycan (b).	72
4.3	Comparison of the two glycan structures identified from the same spectrum by GlycoNovoDB and <i>de novo</i> sequencing algorithm respectively. Glycan (a) identified by GlycoNovoDB has higher score than glycan (b) which is reported by <i>de novo</i> sequencing method.	76
4.4	Another example of two glycan structures identified from the same spectrum by GlycoNovoDB and <i>de novo</i> sequencing algorithm respectively. Glycan (a) identified by GlycoNovoDB has higher score than glycan (b) which is reported by <i>de novo</i> sequencing method.	77
4.5	Comparison of four pairs of glycan structures in the case that glycans identified by <i>de novo</i> sequencing algorithm are ranked top according to the Function 4.10. The first glycan in each row (<i>i.e.</i> a1, b1, c1, d1) is reported by <i>de novo</i> sequencing method, while the second glycan in each row (<i>i.e.</i> a2, b2, c2, d2) is ranked top in the results reported by GlycoNovoDB.	78

List of Tables

2.1	Common monosaccharides	7
2.2	Common amino acid residues	9
2.3	Comparison of the typical performance characteristics of several commonly used mass analyzers.	14
3.1	Mass representations for the ion fragments of the N-linked glycan core structure	37
3.2	Performance of our algorithm compared with GlycoMasterDB	50
3.3	Performance of improved method compared with previous <i>de novo</i> sequencing method	54
4.1	Performance of <i>de novo</i> sequencing method and GlycoNovoDB compared with GlycoMaster DB	70

List of Abbreviations and Terminologies

CID	Collision-induced dissociation
ECD	Electron-capture dissociation
ETD	Electron-transfer dissociation
HCD	Higher-energy collisional dissociation
MS/MS	Tandem mass spectrometry
m/z	Mass-to-charge ratio
PTM	Post-translational modification
glycan	Carbohydrate portion of a glycoconjugate, such as a glycoprotein, glycolipid, or a proteoglycan
glycopeptide	Peptides that contain carbohydrate moieties (glycans) covalently attached to the side chains of the amino acid residues that constitute the peptide
glycoproteomics	A branch of proteomics that identifies, catalogs, and characterizes proteins containing carbohydrates as a post-translational modification
glycosylation	Reaction in which a carbohydrate is attached to a hydroxyl or other functional group of another molecule

List of Algorithms

1	Peptide Mass Inference	38
2	Glycan Candidates Generation by <i>De Novo</i> Sequencing	43
3	String Representation for A Glycan Tree	46
4	Tree Isomorphism Determination	47
5	Filtration of glycan candidates	64

Chapter 1

Introduction

Glycoproteomics is a branch of proteomics that identifies and characterizes proteins containing carbohydrates as a post-translational modification (PTM). As distinct from proteomics, it focuses on the study of the glycosylation of proteins. Glycosylation is one of the most abundant and essential PTMs of proteins. It is frequently observed and over half of eukaryotic proteins in nature are estimated to be glycoproteins [1]. Glycoproteins are involved in a variety of biological processes such as recognition between cell types and immune response to pathogen infections [2]. Research has reported that abnormal glycosylation can lead to serious physiological disorders [3]. Moreover, glycoprotein analysis plays vital roles in drug preparation, such as the design and production of antibodies with selected specificity and function [4]. Additionally, unlike other simple PTMs, which have fixed mass change, glycosylation is much more complex due to its variety of compositions in biological systems and different linkages to proteins [5]. As a consequence, structure analysis in glycoproteomics is more complicated compared with sequencing analysis in conventional proteomics. Hence, developing computational methods for glycoprotein identification is becoming increasingly demanding and remains challenging in glycoproteomics research.

During the past decade, tandem mass spectrometry (MS/MS) has gradually served as a

popular technique for protein sequence identification and quantification [6]. It enables high-throughput protein analysis with high sensitivity and accuracy. In glycoproteomics research, tandem mass spectrometry has also been widely used for the characterization of glycans and glycoproteins in complex biological samples. The peptide with its attached glycan can be studied as a single entity, which provides a comprehensive view of protein glycosylation. Alternatively, each glycan can also be considered as a separate unit [7]. Two different experimental strategies will be conducted to analyze glycoproteins in biological samples, depending on whether glycans and peptides are separated or not prior to the mass spectrometry analysis [8]. One obtains glycan and peptide MS/MS separately and the other generates spectra from intact glycopeptides. The latter strategy has the advantage that it can conserve the glycosylation site information. The general workflow in glycoproteomics analysis with tandem mass spectrometry includes glycoprotein or glycopeptide enrichment, protein or peptide separation, tandem mass spectrometric analysis, and bioinformatic data interpretation [9, 10]. A tandem mass spectrometer can rapidly generate a large amount of mass spectral data for a biological sample, which makes manual interpretation of these data time-consuming and challenging. Effective algorithmic solutions that can facilitate automated analysis of the collected spectral data are required.

With various existing fragmentation methods for tandem mass spectrometers, different kinds of output mass spectra have their unique properties. The commonly used fragmentation techniques include collision-induced dissociation (CID), higher-energy collision dissociation (HCD), electron-capture dissociation (ECD), and electron-transfer dissociation (ETD). Different fragmentation mechanisms usually break at different sites of a glycan or a glycopeptide, and tend to generate different types of dominant fragment ions resulting in different spectra for the same glycan or glycopeptide consequently [6]. It is possible to design effective algorithms by taking advantage of the characteristics of these spectral data or combining multiple types of mass spectral data.

Currently, many efforts have been made to develop approaches for automated interpretation of mass spectrometry based glycoproteomics data. One extensively studied method is the database search, which is to find the best matching glycans by searching the glycan database and comparing theoretical mass spectra with the experimental mass spectra. Several published software packages using the database search method include GlycoFragment and GlycoSearchMS [11], GlycoPep DB[12], GlyDB [13], SimGlycan [14], GlycoPeptideSearch (GPS) [15], GlycoFragwork [7], GlycoMaster DB [16], and MAGIC [17]. Another method for interpreting glycoproteomics data is *de novo* sequencing, which is essential to identifying novel or unknown glycans. The computation of *de novo* sequencing does not depend on database knowledge; instead the algorithms construct glycan structures from mass spectra directly. There have been several attempts to characterize glycan structures from MS/MS in the *de novo* manner, such as GLYCH [18], Peptonist [19], GlycoMaster [20], and the ones proposed by Dong *et al* [21], Böcker *et al* [22], and Sun *et al* [23, 24]. Recently, several algorithms and software tools have been developed to use the spectrum library search approach to solve the peptide identification problem [6]. This method is used to identify peptides by matching the experimental spectrum with the library spectrum directly. This is possible because the size of the spectral library containing annotated spectra with validated results is growing. With the increasing number of publicly available glycopeptide and glycan mass spectrometry data, the spectrum library search method can also be applied to glycan or glycopeptide characterization in the future.

The remainder of this thesis is organized as follows,

Chapter 2 introduces the fundamentals of MS/MS based glycoproteomics research, which include biochemical basics for glycoproteins and mass spectrometry technology. Both experimental and computational strategies for glycoproteomics analysis are described to facilitate the understanding of the subsequent research topics.

Chapter 3 mathematically formulates the problem of glycan *de novo* sequencing with tandem mass spectrometry. Additionally, it presents a heuristic algorithm for glycan *de novo* sequencing from HCD MS/MS spectra of N-linked glycopeptides. The algorithm proceeds in a carefully designated pathway to construct the best matched glycan tree structures from MS/MS spectra. The proposed method has been applied to HCD MS/MS spectra and compared with other methods of similar purpose to evaluate its performance.

Chapter 4 presents a new approach for matching input spectra with glycan structures from a glycan structure database by incorporating a *de novo* sequencing assisted ranking scheme. This approach has been implemented as a software tool named GlycoNovoDB, for automated glycan identification from glycopeptide HCD MS/MS. Experimental results have shown that GlycoNovoDB can identify glycans effectively and has better performance than our *de novo* sequencing algorithm proposed in Chapter 3 as well as another software GlycoMaster DB. In order to identify glycans that are in the database with high accuracy as well as provide new glycans that are not in the database with confidence, an improved method is proposed by further combining the database search method and *de novo* sequencing method together.

Finally, Chapter 5 briefly summarizes the major contents of this thesis and discusses possible future research.

Chapter 2

Background

2.1 Fundamentals of Glycoproteins

2.1.1 Carbohydrates

Carbohydrates are the most abundant type of organic molecules found in nature. A carbohydrate is a biological molecule consisting of carbon (*C*), hydrogen (*H*) and oxygen (*O*) atoms. Their basic molecular formula is $(CH_2O)_n$, where $n = 3$ or more. Carbohydrates can be generally divided into three groups: monosaccharides, oligosaccharides, and polysaccharides. The monosaccharides are also called simple sugars, and cannot be further hydrolyzed to simpler compounds under mild conditions. Oligosaccharides consist of from two to ten simple sugar molecules. Disaccharides and trisaccharides occur frequently in nature. Oligosaccharides with four to six sugar units are usually bound covalently to other molecules, including glycoproteins [25]. Polysaccharides are polymeric carbohydrate molecules composed of long chains of sugar units. They may be either in linear form or highly branched.

Monosaccharides are the building blocks of carbohydrates. Monosaccharides can be classified by the number of carbon atoms they contain: triose (3 carbons), tetrose (4 carbons),

pentose (5 carbons), hexose (6 carbons) etc. Monosaccharides may exist in linear form or ring form, as shown in Figure 2.1. The monosaccharide shown in this figure is a mannose. The ring form of a mannose is a ring structure consisting of six covalent bonds. By convention, the carbon atoms are numbered from 1 to x along the backbone, starting from the end that is closest to the C=O group.

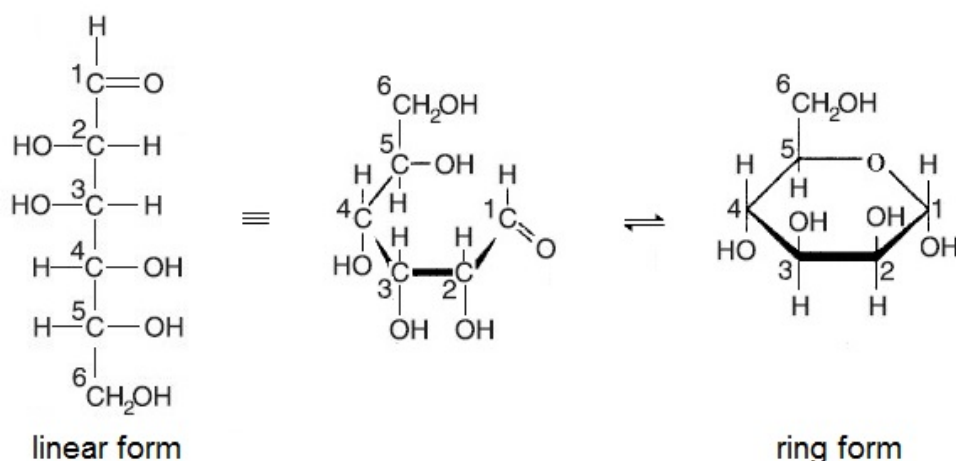


Figure 2.1: The linear form and ring form of a monosaccharide (mannose).

There are numerous types of monosaccharides in nature. Common monosaccharides found in animal oligosaccharides, which are considered in this study, are listed in Table 2.1. It can be observed that some of the monosaccharides are epimers, which means two monosaccharides differ only in their configurations rather than their mass. If two monosaccharides have the same mass, they are equivalent in mass spectrometry. Epimers are not taken into consideration in this study, because they can hardly be distinguished in an MS/MS spectrum. Therefore, in the following thesis, abbreviations are used for monosaccharides with the same formulas. For instance, N-Acetylglucosamine and N-Acetylgalactosamine have the same formula $C_8H_{15}NO_6$, and then their abbreviation HexNAc is used to represent them.

Two monosaccharides can combine via condensation reactions and form a glycosidic bond. The condensation reaction happened between a hemiacetal group (*i.e.* the C_1 group)

Table 2.1: Common monosaccharides

Monosaccharide	Abbreviation	Composition	Monoisotopic mass		Symbol
			Intact	Residue	
Xylose	Xyl	$C_5H_{10}O_5$	150.0528	132.0423	★
Fucose	Fuc	$C_6H_{12}O_5$	164.0685	146.0579	▲
Glucose Mannose Galactose	Hex	$C_6H_{12}O_6$	180.0634	162.0528	○
N-Acetylglucosamine N-Acetylgalactosamine	HexNAc	$C_8H_{15}NO_6$	221.0899	203.0794	□
N-Acetylneuraminic acid	NeuAc	$C_{11}H_{19}NO_9$	309.1060	291.0954	◆
N-Glycolylneuraminic acid	NeuGc	$C_{11}H_{19}NO_{10}$	325.1009	307.0903	◇

of one monosaccharide (or a molecule derived from a saccharide) and the hydroxyl group of the other. During the reaction, an OH group is removed from one of the sugars and an H from the other and an H_2O molecule is formed. The new bond that is formed to combine the sugars together is known as a glycosidic bond, or glycosidic linkage. Figure 2.2 shows two glucose molecules combine with each other to form a maltose and release a water. Depending on which hydroxyl group participates in the reaction, there are four possible types of glycosidic bonds, which are 1-2, 1-3, 1-4, and 1-6. The numbers in the notation represent the numbering of carbon atom in the hemiacetal group and the hydroxyl group, respectively. As shown in Figure 2.2, the C_1 of one molecule reacts with the C_4 of the other molecule, and this forms a 1-4 glycosidic bond.

Figure 2.3(a) shows an example of tetrasaccharide consisting of four glucose molecules. The glycosidic bonds shown in the tetrasaccharide are two 1-4 bonds and one 1-6 bond respectively. The structure of an oligosaccharide can be represented as a tree structure, as shown in

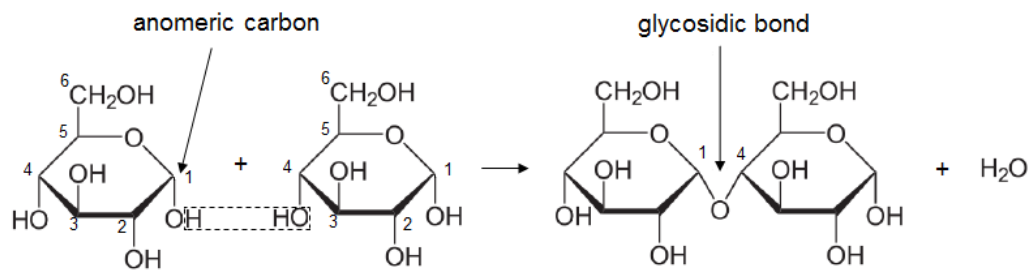


Figure 2.2: Two glucose molecules combine to form a maltose and release a water. The hemi-acetal carbon atom (C_1) is called anomeric carbon.

Figure 2.3(b). In the tree representation, each monosaccharide is represented as a node and we use a symbol to denote it, meanwhile each glycosidic bond is represented as an edge. Since there are at most five linkages for one monosaccharide, the degree of a glycan tree is bounded by four.

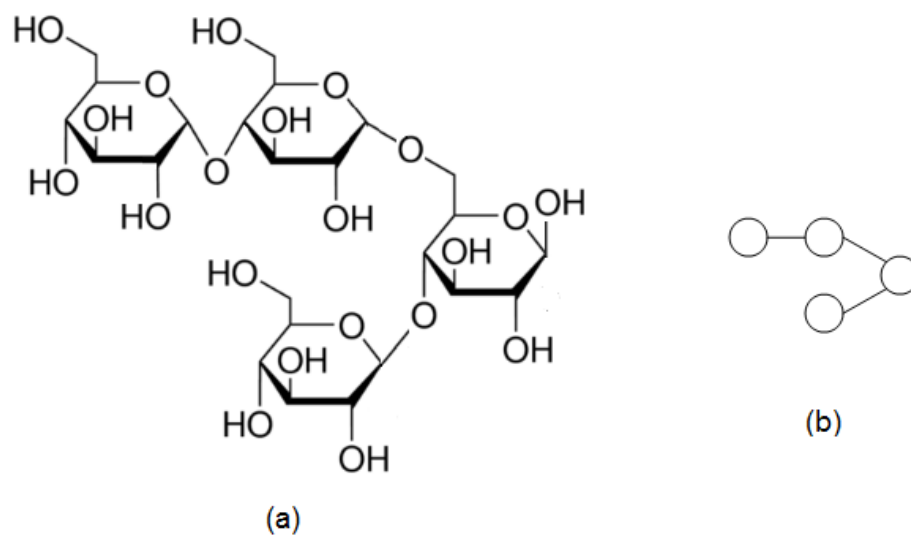


Figure 2.3: An example of tetrasaccharide and its tree representation.

Table 2.2: Common amino acid residues

Name	3-letter Symbol	1-letter Symbol	Mono Mass	Avg. Mass	Residue Composition
Alanine	Ala	A	71.03711	71.08	C ₃ H ₅ NO
Arginine	Arg	R	156.10111	156.2	C ₆ H ₁₂ N ₄ O
Asparagine	Asn	N	114.04293	114.1	C ₄ H ₆ N ₂ O ₂
Aspartic Acid	Asp	D	115.02694	115.1	C ₄ H ₅ NO ₃
Cysteine	Cys	C	103.00919	103.1	C ₃ H ₅ NOS
Glutamic Acid	Glu	E	129.04259	129.1	C ₅ H ₇ NO ₃
Glutamine	Gln	Q	128.05858	128.1	C ₅ H ₈ N ₂ O ₂
Glycine	Gly	G	57.02146	57.05	C ₂ H ₃ NO
Histidine	His	H	137.05891	137.1	C ₆ H ₇ N ₃ O
Isoleucine	Ile	I	113.08406	113.2	C ₆ H ₁₁ NO
Leucine	Leu	L	113.08406	113.2	C ₆ H ₁₁ NO
Lysine	Lys	K	128.09496	128.2	C ₆ H ₁₂ N ₂ O
Methionine	Met	M	131.04049	131.2	C ₅ H ₉ NOS
Phenylalanine	Phe	F	147.06841	147.2	C ₉ H ₉ NO
Proline	Pro	P	97.05276	97.12	C ₅ H ₇ NO
Serline	Ser	S	87.03203	87.08	C ₃ H ₅ NO ₂
Threonine	Thr	T	101.04768	101.1	C ₄ H ₇ NO ₂
Tryptophan	Trp	W	186.07931	186.2	C ₁₁ H ₁₀ N ₂ O
Tyrosine	Tyr	Y	163.06333	163.2	C ₉ H ₉ NO ₂
Valine	Val	V	99.06841	99.13	C ₅ H ₉ NO

2.1.2 Amino Acids

Amino acids are molecules containing an amine group ($-NH_2$), a carboxylic acid group ($-COOH$) and a side chain (R group) that specific to each amino acid. Amino acids are building blocks of proteins. The amino and carboxyl groups of amino acids can react in a head-to-tail fashion, eliminating a water molecule and forming covalent amide linkage, which is typically referred to peptide bone in peptides or proteins. There are 20 common amino acids commonly found in proteins. Table 2.2 lists the name, mass and composition information of the 20 common amino acids.

2.1.3 Glycans and Glycoproteins

Carbohydrates are covalently linked with a variety of other molecules, such as lipid molecules, which are common components of biological membranes. Proteins that are covalently linked to carbohydrates are called glycoproteins. Glycoproteins, together with glycolipids, are called glycoconjugates, and they are important components of cell walls and extracellular structures in plants, animals, and bacteria [25]. Besides, they also serve in a variety of processes involving recognition between cell types or recognition of cellular structures by other molecules.

Glycans usually refer to carbohydrate chains attached to glycoproteins or glycolipids. Glycosylation occurs when a glycan is linked to a protein at specific amino acid residues. Two different types of glycoproteins are commonly observed: N-linked glycoproteins and O-linked glycoproteins, as shown in Figure 2.4. N-linked glycoproteins involve the attachment of carbohydrate groups to the amide nitrogen of an asparagine residue in the peptide chain [25]. In O-linked glycoproteins, glycans are attached to proteins via the hydroxyl group of a serine or threonine residue [25]. Glycans in N-linked glycoproteins are called N-linked glycans. O-linked glycans refer to those in O-linked glycoproteins.

Analysis of protein sequence database has revealed that in most cases N-linked glycans are attached to proteins via a sequence motif Asn-Xxx-Ser or Asn-Xxx-Thr, where Xxx denotes any amino acid except proline [26]. The consensus tripeptide Asn-Xxx-Cys is also possible in N-linked glycoproteins but less frequently observed [27]. These motifs provide good information for the analysis of glycosylation site and the sequencing of glycan-linked peptide. Research has shown that most mammalian N-linked glycans share a common core structure composed of two N-acetylglucosamine residues linked to a branched mannose triad. This core structure is attached to asparagine residue in the peptide sequence via N-acetylglucosamine. Other sugar units may be attached to each of the mannose residues of this branched core. The resulting structures fall into three main categories of N-linked glycoforms: high mannose,

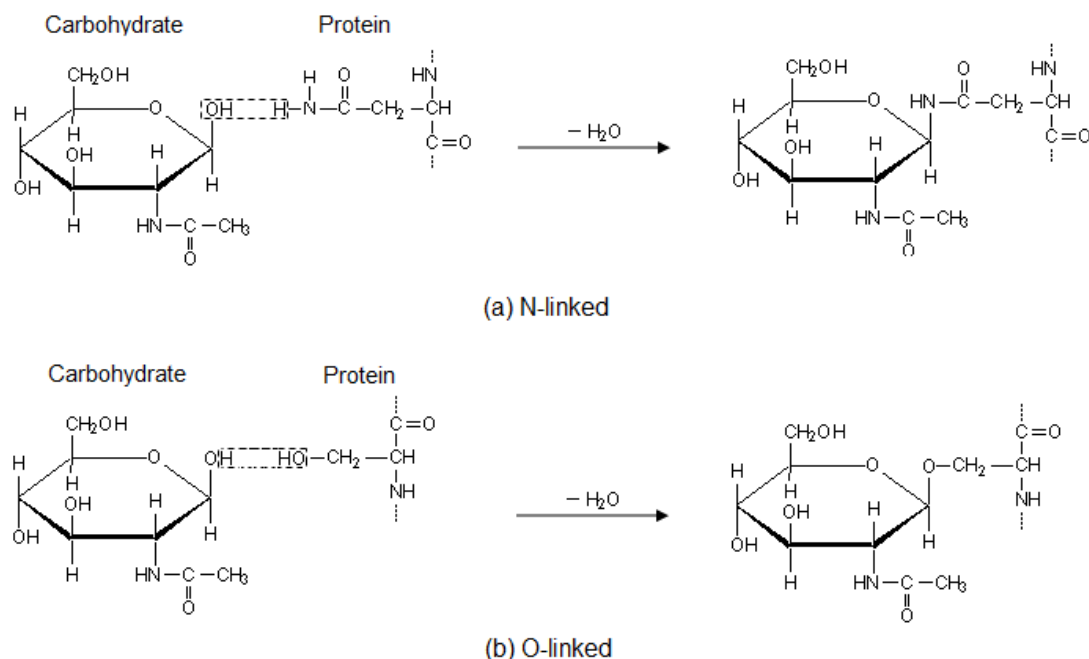


Figure 2.4: N-linked and O-linked glycoproteins.

complex, and hybrid, depending on what types of other sugars are attached to this core.

Unlike N-linked glycoproteins, which have common core structure and motif, O-linked glycoproteins have more varied core structures and peptide sequences. The work of this study mainly focuses on the analysis of N-linked glycoproteins.

2.2 Mass Spectrometry Technology

Mass spectrometry (MS) is an analytical technique that determines the composition of a sample by measuring the mass-to-charge ratio (m/z) of the ionized analytes. Mass spectrometry has both qualitative and quantitative uses, which include identifying the composition and structures of unknown compounds [28–30], and quantifying the amount of a compound in a sample [31–34]. Nowadays, MS is commonly used in analytical laboratories where need to analyze physi-

cal, chemical, or biological properties of a wide variety of compounds. In proteomics, MS has gradually become the method of choice for analysis of complex protein samples.

2.2.1 Mass Spectrometers and Configuration

In mass spectrometry, the molecules are ionized and the mass-to-charge ratio (m/z) of the ions are measured, rather than measure the mass of a molecule directly [6]. A mass spectrometer typically consists of three components: an ion source, a mass analyzer that measures the m/z of the ionized analytes, and a detector that registers the number of ions at each m/z value [35]. In a typical MS procedure, molecules of interest are first ionized in the ionizer. The ions are then separated according to their different mass-to-charge ratio in the mass analyzer. Finally the separated ions are detected by a mechanism capable of detecting charged particles. Results are displayed as mass spectra, each of which consists of a list of peaks. Each peak is represented by its m/z value and the relative abundance (*i.e.* intensity) of detected ions.

Each of the three major components of a mass spectrometer can be implemented based on different technologies, generating mass spectral data with different properties. Two techniques commonly used to ionize the molecules for mass spectrometric analysis are matrix-assisted laser desorption/ionization (MALDI) [36, 37] and electrospray ionization (ESI) [38, 39]. The main difference between these two ionization techniques is that MALDI produces singly charged ions ($z = 1$) while ESI can produce singly and multiply charged ions (mainly $z \geq 1$). ESI has the advantage that a large molecule can still be detected because its ions can fall into the m/z range of a mass spectrometer when the charge state $z \geq 1$. Thus, MALDI-MS is normally used to analyze relatively simple peptide mixtures, whereas integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex samples [35].

Mass analyzer, among the three components, is central to the mass spectrometry technology. In proteomics, three primary parameters of mass analyzer are considered, which include mass resolving power, mass accuracy, and mass range. Mass analyzers separate ions based on their m/z . The mass resolving power evaluates how well the separations can be performed and measured. IUPAC [40] defines resolving power in mass spectrometry is $M/\Delta M$, where ΔM is the minimum peak separation, and M refers to the mass of the (second) peak. ΔM measures the minimum peak separation in mass spectrometry which can be defined in different ways. Two widely used definitions are the valley definition and the peak width definition. The valley definition defines ΔM as the closest spacing between two singly charged ion signals of equal height with the valley between them less than a specified fraction of the height of either peak. Typical values of the fraction are 5%, 10%, or 50%. In the peak width definition, ΔM refers to the width of a single peak at a height which is a specified fraction of the maximum peak height. In practice, the value of 50% is frequently used, and is termed the “full width at half maximum” (FWHM). Figure 2.5 illustrates the two definitions.

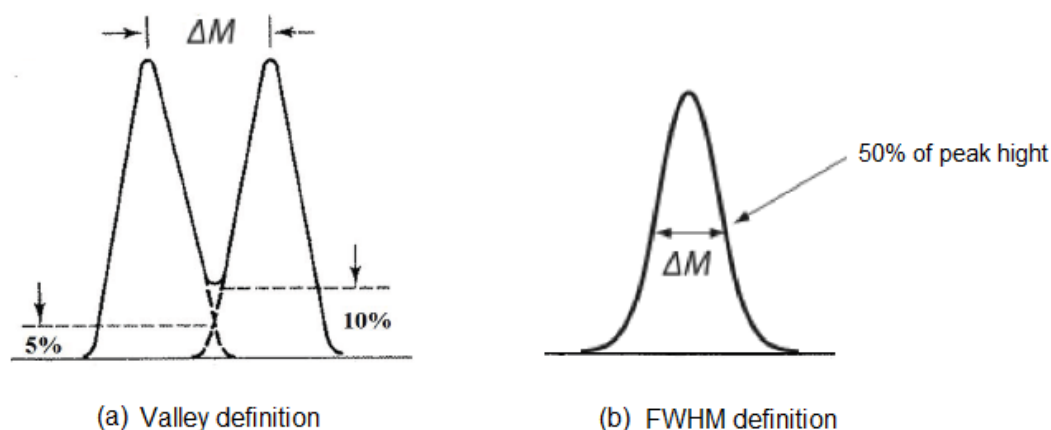


Figure 2.5: Methods of calculating mass resolving power: (a) calculation via 5% and 10% valley definition. (b) calculation via full width at half maximum (FWHM) definition.

Mass accuracy refers to the difference between the true m/z and the measured m/z of a given ion divided by the true m/z of the ion. It is usually measured by the terms of parts per mil-

lion (ppm). Higher mass accuracy can increase the degree of confidence for peak assignments, because increasing in mass accuracy will enhance the possibility of uniquely identifying the elemental compositions of observed ions [40]. The mass range is the range of m/z over which a mass analyzer can operate to record a mass spectrum. The unit of mass range is Dalton (Da). One Da is $\frac{1}{12}$ of the mass of a carbon atom (^{12}C), and is approximately the mass of a hydrogen atom.

There are several basic types of mass analyzer commonly used in proteomics research. These are quadrupole [41], ion trap (quadrupole ion trap, QIT [42], linear ion trap, LIT or LTQ [43]), time-of-flight (TOF) [44], Fourier transform ion cyclotron resonance (FTICR) [45], and orbitrap [46]. Different types of mass analyzer are very different in design and performance, in terms of sensitivity, accuracy, mass range, and other properties [40, 47]. These analyzers can be used alone or put together in tandem mass spectrometry to utilize each advantages. Table 2.3 [47, 48] summarized the performance of each mass analyzer.

Table 2.3: Comparison of the typical performance characteristics of several commonly used mass analyzers.

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

The final element of a mass spectrometer is the ion detector. After the ions are separated by mass analyzer, the detector records the current signal produced or the charge induced when an ion passes by or hits the metal surface of the detector. Usually the mass spectrometer is connected to a computer with software that analyze the data provided by ion detector and produce mass spectra. A mass spectrum is an intensity vs. m/z plot representing the distribution of ions

by m/z in a sample. Each spectrum consists of a list of peaks. Figure 2.6 illustrates an example of a mass spectrum. The x-axis of the mass spectrum represents mass-to-charge ratio (m/z), and y-axis represents signal intensity of the ions. Each peak in the spectrum represents the ions with the same m/z value, and the intensity of the peak reflects the number of ions detected by the detector at the m/z . However, the abundances ratio between two different molecules can not be regarded as the ratio of their peak intensities directly, because not all molecules in the sample are measured with the same efficiency. Due to the charge competition [49] and detectability of different molecules, some molecules may produce much lower intensity peaks than other molecules even if they have the same abundance level in the sample.

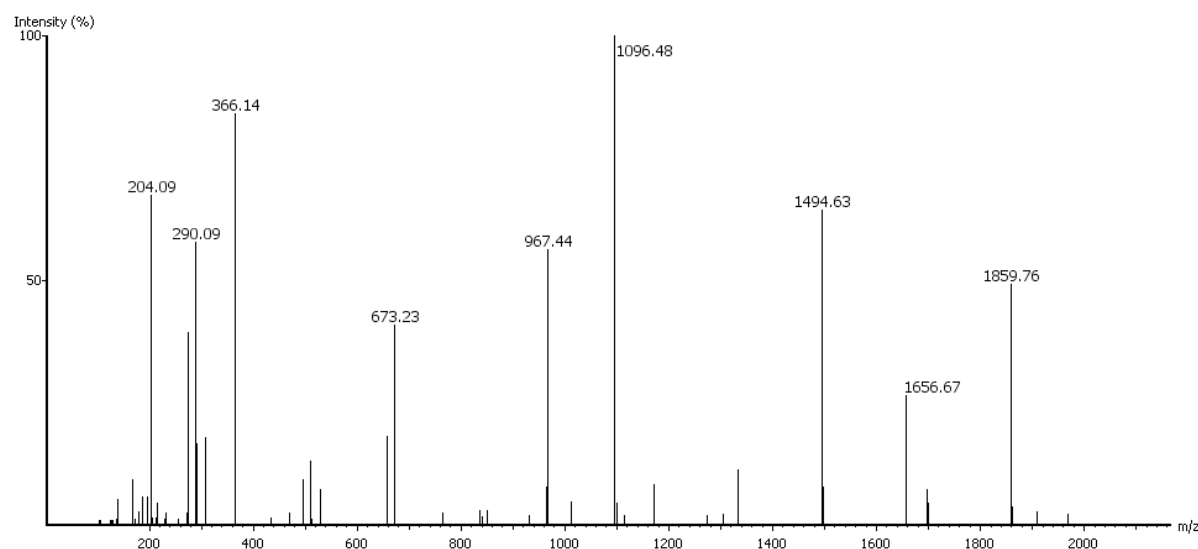


Figure 2.6: An example of a visualized mass spectrum. The x-axis of the mass spectrum represents mass-to-charge ratio (m/z), and y-axis represents signal intensity of the ions.

2.2.2 Tandem Mass Spectrometry

Tandem mass spectrometry (MS/MS) is a reliable tool to precisely identify and characterize the structures of molecules, because it can provide more information about the molecules of interest than traditional mass spectrometry. A tandem mass spectrometer has two mass analyzers, or two sequential analyses in the same analyzer. The first mass analyzer selects ions at a

certain m/z window. The selected ion is called precursor ion or the parent ion, which is then fragmented by some fragmentation methods and yields product ions. The second mass analyzer measures the product ions as usual to form tandem mass spectrum. Peaks in the tandem mass spectrum represent a set of fragment ions generated from the dissociation of a selected molecule. Figure 2.7 shows the schematic of a typical tandem mass spectrometry. Two types of mass spectra are generated in an MS/MS experiment: survey scan (or MS spectrum) and tandem mass spectrum (or MS/MS spectrum).

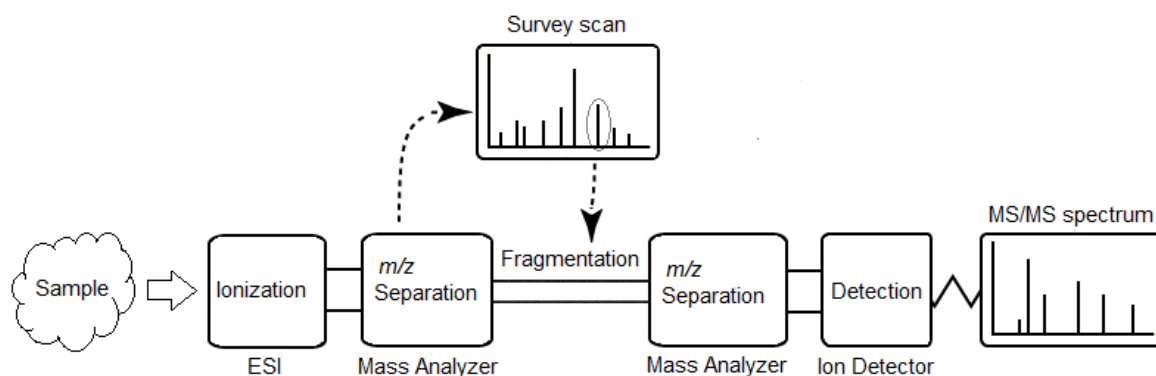


Figure 2.7: Schematic of tandem mass spectrometry. The sample is ionized in the mass ionizer first, and then analyzed by the first mass analyzer. Ions of interest (*i.e.* precursor ion) are selected from the generated survey scan and fragmented into product ions, which are analyzed by the second mass analyzer. Finally, MS/MS spectrum is produced.

In tandem mass spectrometry, fragmentation is an important step, which can help to generate the structural information of a molecule by ion dissociation. The most widely applied fragmentation method for proteome identification and quantification is collision-induced dissociation (CID) [50], or collisionally activated dissociation (CAD). Under CID condition, the molecular ions undergo one or more collisions by interactions with neutral gas molecules, contributing to vibrational energy which results in bond breakage and the fragmentation of the molecular ion into smaller fragments. In general, CID is more effective for small, low-charged peptides. Complementary to CID fragmentation, electron-transfer dissociation (ETD) [51], or electron-capture dissociation (ECD) induces cleavage of charged molecules by transferring electrons to them. Ideally, ETD can provide both the sequence information and the localization

of the modification sites for the peptides with PTMs. Another frequently used fragmentation method is high-energy collision dissociation (HCD) [52]. HCD fragmentation method is featured with higher activation energy and shorter activation time comparing the traditional CID fragmentation method. Different fragmentation methods result in ion dissociation occurring at different sites and can generate different types of dominant fragment ions. We will discuss the fragmentation patterns of glycopeptide in the next section.

2.3 Mass Spectrometry Based Glycoproteomics Analysis

In glycoproteomics analysis, mass spectrometry has become a powerful tool because of its high sensitivity and throughput. More specifically, mass spectrometry has been widely used to identify glycoproteins, to evaluate glycosylation sites, and to elucidate glycan structures [53–55]. Although both MALDI and ESI are capable of recording spectra of intact glycoproteins, currently individual glycoforms can only be resolved from small proteins containing a limited number of glycans, preferably attached to a single site [55]. Therefore, the top-down approach for glycoprotein characterization in a complex sample is still challenging. The most widely used methods are based on characterizing glycopeptides generated by the digestion of glycoproteins, followed by the analysis which based on either intact glycopeptides or deglycosylated glycopeptides.

2.3.1 General Strategies

Depending on whether glycans and peptides are separated or not before the mass spectrometry analysis, there are generally two different strategies to analyze glycoproteins in biological samples [8, 56, 57]. Figure 2.8 illustrates these two strategies for an integrated glycoproteomics analysis. In one workflow, deglycosylation is applied to the glycoproteins in the biological

samples first, in order to obtain glycans and deglycosylated glycopeptides respectively. The mass spectrometry analysis in the following step will be conducted on the obtained glycans and peptides separately. Finally the collected mass spectra will be used for glycan and peptide identification. In the other workflow, glycoproteins are digested into glycopeptides by trypsin first, and then the resulting intact glycopeptides are sent to mass spectrometer to produce mass spectra. The final step for this strategy is the same as the previous method, which is to characterize glycans from mass spectral data.

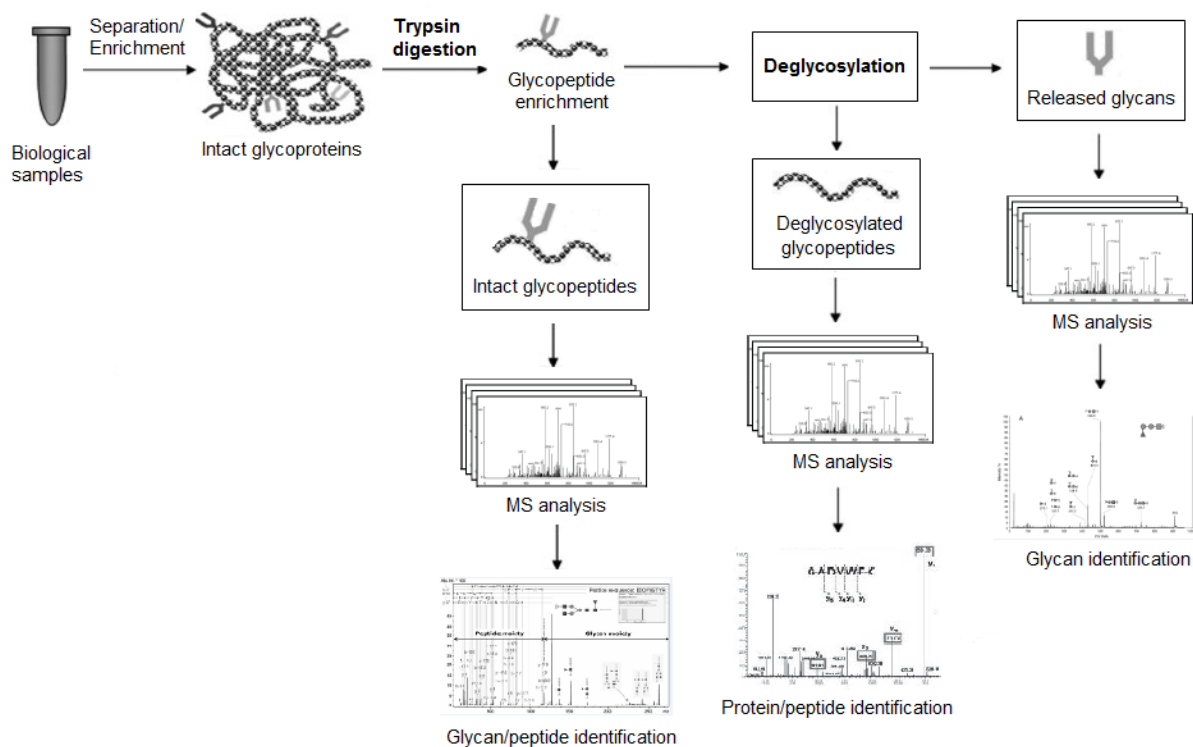


Figure 2.8: The strategies of mass spectrometry based glycoproteomics analysis. One strategy is to isolate glycans from glycopeptides and then generate mass spectra of the released glycans and deglycosylated peptides separately. The other strategy produces mass spectra directly from intact glycopeptides.

Interpreting the mass spectral data of released glycans is a straightforward way to identify glycan structures. However, it is difficult for assigning glycans as information about the sites of glycan attachment cannot be inferred from MS results automatically. Instead, analysis of the

data from intact glycopeptides has the advantage that the glycosylation sites can be reserved and used to identify attached peptide sequence. Experiments have been reported to characterize intact glycopeptides by interpreting tandem mass spectra of glycopeptides [8, 53, 58].

2.3.2 Fragmentation Patterns

In tandem mass spectrometry experiments, selected glycopeptide precursors will be further fragmented before second mass analyzer analysis. The two components, peptide and glycan, of the glycopeptide will undergo fragmentation. Peptide is a linear polymer of amino acids. During the process of fragmentation, the peptide ion can fragment at three different sites along the amino acid backbone, as shown in Figure 2.9. The nomenclature for fragment ions was first proposed by Roepstorff and Fohlman [59], and subsequently modified by Johnson et al. [60]. Fragments will only be detected if they carry at least one charge. As shown in the figure, if the charge is retained on the N terminal fragment, the ions are categorized as either *a*, *b* or *c*. If the charge is retained on the C terminal, the ion type is either *x*, *y* or *z*. The subscript indicates the number of residues in the fragment. The mass value of each peptide fragment ion provides peptide structural information.

Similarly, the glycan moiety of a glycopeptide will be fragmented into different fragment ions. According to cleavages at different sites of glycosidic bonds, there are six basic types of ions: *B/Y*-ions, *C/Z*-ions, and *A/X*-ions, as shown in Figure 2.10. Fragment ions that contain a non-reducing terminus (the monosaccharide residue in acetal form) are labelled with letters A, B, C, and those that contain the reducing end (the monosaccharide residue with hemiacetal functionality) of the glycan are labelled with letters from the end of the alphabet (X, Y, Z); subscripts indicates the cleaved ions [61]. X-, Y- and Z-ions retain peptide units. The A-ions and X-ions are produced by cross-ring cleavages, and are labelled by assigning each ring bond a number and counting clockwise. If two or more glycosidic bond cleavages happen at the

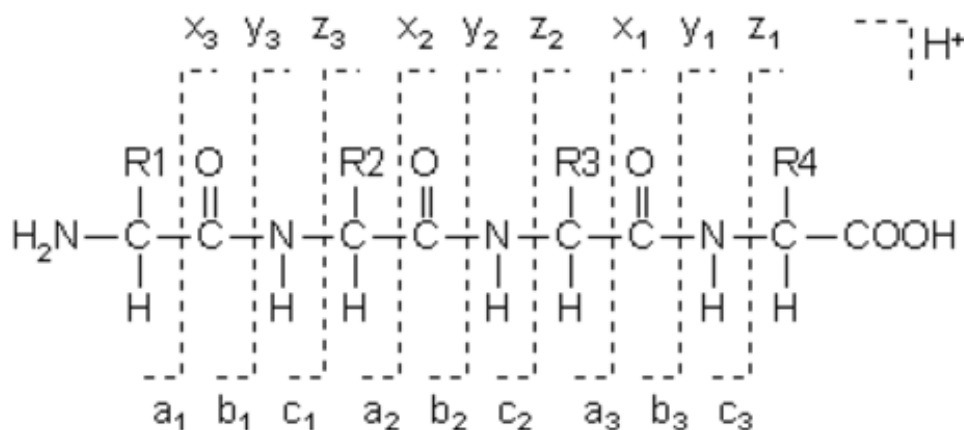


Figure 2.9: Examples of fragmented ions of a peptide. This peptide consists of four amino acids. Six basic types of fragment ions, including *a*-, *b*-, *c*-, *x*-, *y*-, and *z*-ions, generated by cleavages on the peptide backbone.

same time, internal fragment ions will be produced.

In MS/MS, different fragmentation approaches result in ions dissociation at different glycosidic bonds or peptide backbone, and are consequently featured with different types of dominant fragment ions. In recent research, several different dissociation methods have been reported for glycoproteomics analysis, such as CID, HCD, and ECD/ETD. Each of these fragmentation methods has its unique characteristics and can be chosen according to different requirements. CID and HCD usually break the glycosidic bonds of glycopeptide and yield *B*-ions and *Y*-ions. HCD spectra are featured with a predominance of *Y*-ions. *B*-ions and *A*-ions and other smaller species produced by further fragmentations can also be occasionally observed [62]. Besides, HCD can produce fragment ions by breaking the glycosidic bonds but leaving the attached peptide intact. In contrast, ECD/ETD often lead to cleavages at peptide backbone and produce *C*-ions and *Z*-ions. ECD/ETD spectra can be used to determine both the peptide sequence and the glycosylation site because the glycan linked to peptide backbone can be retained intact [63]. Theoretically, both the glycan moiety and peptide backbone undergo

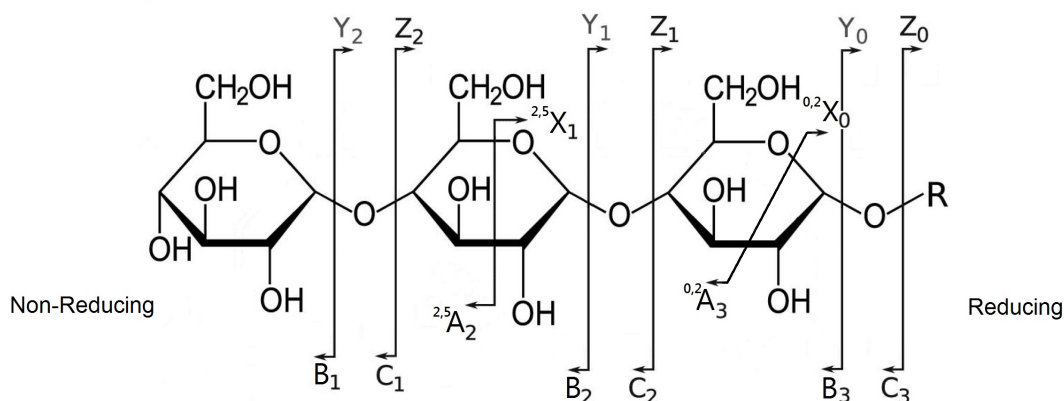


Figure 2.10: Examples of fragmented ions of a glycan. This glycan contains three monosaccharides. A-, B-, C-ions and X-, Y-, Z-ions are shown in the figure, where A/X- ions generated from cross ring cleavages. It is worth noticing that not all the ions are marked in the figure.

fragmentation. CID and HCD techniques are mainly used for glycan structure determination. By controlling the fragmentation energy, we can obtain the tandem mass spectra of glycopeptides in which a majority of peaks are generated from fragmentations upon the glycan moiety.

Combining two complementary fragmentation techniques in MS/MS analysis enables the identification of peptide sequences, glycan structures, and glycosylation sites. For instance, CID/ETD enables the elucidation of glycosylation sites by maintaining the glycan-peptide linkage [64]; HCD/ETD enables the identification of glycan structure and peptide backbone, allowing glycopeptide identification [16, 65]. It has become increasingly popular and promising to use more than one types of fragmentation approaches in proteomics research.

2.4 Computational Approaches for Interpreting MS/MS Spectra of Glycopeptides

In mass spectrometry based glycoproteomics, researchers identify glycan structures, attachment sites and glycosylation linkages by interpreting MS/MS spectra produced in the experi-

ment. The process of MS/MS interpretation and glycan reconstruction used to be done manually by biochemists. However, with the rapid increase of mass spectrometry data collected in wet-lab experiment in recent years, it is impractical for researchers to interpret the mass spectral data manually. Therefore, it is necessary to develop automated approaches to do glycoprotein profiling based on mass spectrometry. Until now, extensive research has been made for mass spectral interpretation. Generally, the computational approaches for automated glycan or glycopeptide identification from mass spectra fall into two categories: database search and glycan/glycopeptide *de novo* sequencing.

2.4.1 Database Search

Database search method is to search the glycan database to find the best matching glycans by comparing theoretical mass spectra with the experimental ones. This kind of approaches require the assistance of glycan database which is supposed to contain all the target glycans. Although the protein sequence database commonly used in proteomics research seldom record the glycan structure information for glycosylated proteins [9, 16], databases for isolated glycan structures have become available, including CCSD/CarbBank [66], SweetDB [67], GLYCO-SCIENCES.de [68], EUROCarbDB [69], GlycoSuiteDB [70, 71] and GlycomeDB [72, 73]. By far, several software packages have been developed based on this strategy.

GlycoFragment [74] can calculate all the theoretically possible MS-relevant fragment ions of the glycans and use them to annotate an MS/MS spectrum. GlycoSearchMS [11] takes the peak values of the experimental mass spectra as an input and searches for matches with the calculated fragments of all glycan structures contained in the SweetDB database. GlycosidIQ [75] was developed for computational interpretation of oligosaccharide mass spectrometric fragmentation based on matching experimental data with theoretically fragmented oligosaccharides generated from the database GlycoSuiteDB. GlycoWorkBench [76] is a soft-

ware tool developed by the EUROCarbDB initiative to provide support for the manual interpretation of MS data. It evaluates a set of structures proposed by the user via matching the list of peaks derived from the spectrum against the corresponding theoretical list of fragment masses. These tools are designed for the interpretation of mass spectrometry data generated from released glycans instead of intact glycopeptides. Since the glycan and peptide are separated before mass analysis, the information regarding the sites of glycan attachment cannot be retrieved. The process of identifying deglycosylated peptide sequences is the same as that is used in general proteomics research. The software tools and algorithms commonly used for analyzing the MS/MS data of peptides include MOWSE [77], Mascot [29], PEAKS DB [78], SEQUEST [79], X!Tandem [80].

GlycoPeptideSearch (GPS) [15] is a glycan database search program that can identify intact N-glycopeptides from CID MS/MS spectral data. It first computes a short list of peptides with an N-linked glycosylation motif, and then filters the MS/MS spectra with the signature ions at m/z 204 and 366, corresponding to oxonium ions formed by a HexNAc and a disaccharide Hex-HexNAc. Glycans from GlycomeDB that match the putative glycan masses are grouped according to their mass and reported as a single match.

Mayampurath *et al.* proposed a computational framework for identifying intact N-linked glycopeptides, and implemented it in a software tool called GlycoFragwork [7]. The tool provided a platform for simultaneous scoring of fragmentation spectra of glycopeptides using different methods, which include CID, HCD, and ETD. In the framework, individual scoring schemes are applied to each fragmentation type utilizing information from a glycan and peptide database, and an empirical false-discovery rate estimation method, based on a target-decoy search approach is derived for the confidently assignment of these glycopeptides.

Two more recent tools, GlycoMasterDB [16] and MAGIC [17], both are designed for automated identification of intact N-linked glycopeptides from HCD MS/MS spectra. Glyco-

MasterDB simultaneously searches a protein sequence database and a glycan structure database extracted from GlycomeDB to find the best pair of peptide and glycan for each input spectrum. The software tool analyzes the data in three steps, including filtration of glycopeptide spectra, glycan assignment, and peptide identification. The first two steps for glycan identification are based on HCD spectra. The third step for peptide sequence determination uses either ETD spectral (if available) or the calculated mass values of the peptides that bear the glycans and contain the motif of N-glycopeptide. MAGIC is a glycopeptide identification platform that can characterize N-linked glycoproteomics using beam-type CID data sets without prior known protein and glycan information. It first uses an algorithm to detect a triplet pattern for accurate Y1-ion identification. Then on the basis of the detected peptide mass, it generates *in silico* peptide MS/MS spectra by assigning precursor mass and removing all of the *B*- and *Y*-ions for database search. Finally, the glycan composition is determined based on the glycan mass and all of the detected *B*- and *Y*-ions.

2.4.2 Glycan *De Novo* Sequencing

The computation of glycan *de novo* sequencing does not rely on glycan database knowledge, instead the algorithms construct glycan structures from mass spectra directly. Therefore, *de novo* sequencing method is essential to identify novel or unknown glycans. Besides, in the situation that the target glycan is not included in the glycan database, this method is often used. Previously, such method was often performed manually by biochemists. Recently, several computational approaches have been proposed to meet the need of higher throughput.

STAT [81] is web-based computational program for saccharide topology analysis. It first uses “knapsack algorithm” to compute all possible combinations of the selected monosaccharides that sum to the adjusted total mass. Besides, the possible compositions of each product ion mass input from tandem mass spectra are computed by essentially the same algorithm. Fi-

nally, STAT computes all of the possible topologies that contain all of the substructures implied by the product ions and rates the possible structures based on the likelihood that it is the correct sequence. STAT can support calculations for structures of up to 10 monosaccharide units.

StrOligo [82] is a computer program for automated interpretation of tandem MS spectra of complex N-linked glycans from mammals. The algorithm first builds a relationship tree by examining each pair of peaks and checking whether the differences in m/z correspond to the losses of one or two known monosaccharides. Then it determines the most probable composition by testing all possible combinations of monosaccharide residues summing up to the precursor m/z value and scores them based on the degree of agreement with the relationship tree. Subsequently, the potential compositions are assigned to a limited set of structures based on the constraints deduced empirically from glycan structures observed in mammals. However, this algorithm has limited ability to identify hybrid and high-mannose N-linked oligosaccharides.

Tang *et al.* proposed a dynamic programming algorithm GLYCH [18] to determine oligosaccharide structures from tandem mass spectra. This program takes cross-ring ions resulting from interval cleavages into consideration when scoring possible glycan structures, which is not incorporated into other programs. The algorithm consists of three steps. First, a scoring scheme is developed to identify potential bond linkages between monosaccharides based on the appearance pattern of cross-ring ions. Next, a dynamic programming algorithm is used to determine the most possible glycan structures from the mass spectra. Finally, a re-evaluation scheme is implemented to re-rank the oligosaccharides generated by taking into account the double fragmentation ions. One limitation of GLYCH algorithm is that it prefers linear structure to branching structure.

In [20], glycan structure *de novo* sequencing is defined as the problem of finding a glycan tree structure T such that the mass of T is equal to a given value (precursor ion mass subtract

peptide mass and a water), and summation of the scoring function calculated according to the peaks in MS/MS is maximized. The time complexity varies according to whether the peaks can be used repeatedly, because several different fragment ions of glycan tree may have the same mass value and produce the same peak. It has been proved that glycan *de novo* sequencing is an NP-hard problem, under the condition that each peak in the spectrum could be used only once. GlycoMaster [20] uses heuristic programming technique to compute the best possible structure among all possible monosaccharide combinations. It first generates many acceptable small subtrees, and then join them together in an iterative process to obtain larger suboptimal subtrees until the desired precursor mass is reached.

In [22], Böcker *et al.* presented an exact algorithm based on fixed-parameter algorithm attempting to solve the NP-hard problem. They modify the recurrences proposed in [20] and limit the running time explosion to the number of peaks in the sample spectrum. Since the number of simple fragments of a given glycan is linear to its number of monosaccharides, the algorithm can be maintained a polynomial running time with respect to the mass value of the glycan. Besides, they also incorporate the set of explained peaks into the dynamic programming to avoid multiple peak counting. At the end of this paper, it shows that how to count the number of glycan topologies for a fixed size or a fixed mass and the complexity of their algorithm is given, which indicates that the proposed algorithm is efficient.

Recently, Dong *et al.* [21] introduced a new glycan representation by abstracting a glycan structure as a directed acyclic graph instead of a tree. Based on this representation, they proposed a *de novo* sequencing algorithm to construct the tree structure from MS/MS spectra with logical constraints and some known biosynthesis rules. The algorithm rebuilds the glycan structure in a bottom up fashion iteratively by storing every confirmed substructures as building block and using them to construct larger substructures. The scoring function used in the algorithm allocates different probability for different glycosidic cleavages when the theoretical spectrum is generated, which profile different candidate structures more accurate.

A more recent program glyfon [83] based on machine learning was developed for *de novo* sequencing of glycans from MS/MS spectra. The program builds a suitable model that disallows multiple assignments for each peak of the fragmentation ions of glycans, and implements a solver that employs Lagrangian relaxation with a dynamic programming technique. Then it introduces a machine learning technique called structured support vector machines to optimize scores for the algorithm. Additional constraints for core structures of known glycan types are also implemented in the algorithm. An improvement of this program is investigating a novel scoring scheme which takes intensities of MS/MS spectra into account.

For the purpose of interpreting glycan structures from mass spectra, database search method generally has the ability to obtain more reliable results than *de novo* sequencing method, due to that it is equipped with glycan database containing the structural knowledge of glycans. Database search is expected to become more powerful with the size of glycan database increasing. In the meanwhile, *de novo* sequencing method requires relatively high quality spectra to generate more accurate results. However, with the development of mass spectrometry technology, it has shown promising research prospects in glycan identification. It is unavoidable that we will observe some new glycan structures which are not included in glycan database during proteomics data analysis. As we know, manual interpretation of the glycopeptide spectra is time-consuming and usually unreliable, which indicates that the interpretation of the novel glycan by *de novo* sequencing method is necessary and sometimes the only available way.

Chapter 3

Glycan *De Novo* Sequencing from HCD Spectra

There are generally two strategies for glycoproteomics research depending on whether the glycopeptides are analyzed with the glycan released or kept attached to the peptide. Most of the previously available software tools or algorithms are designed to characterize released glycans only. One disadvantage of this strategy is that it is hard to determine the glycosylation site if there are more than one glycans attached to the glycoprotein or multiple glycoproteins in the sample are digested into glycopeptides. Furthermore, it is also difficult for the sequenced peptide to find the corresponding glycans attached to it. Generally, it is considered to be a better way to identify glycans from the MS/MS spectra generated from intact glycopeptides rather than released glycans.

GlycoMaster [20] is a useful software tool that can identify glycan structures from tandem mass spectra without the knowledge of glycan database. It gradually constructs larger glycan tree structures from previously selected subtrees using a heuristic algorithm. During each round of growing larger trees, the best ranked subtree structures are reported and added in the candidate list for the next round. However, this approach of constructing trees from leaves to root has an unavoidable drawback. The algorithm was initially designed to determine glycan

structures from CID MS/MS of glycopeptides. Due to the limited dynamic range of CID spectrum, the mass values of *Y*-ion fragments attached to the peptide will normally exceed the mass range of the spectrum. Besides, the peaks in the lower mass end of the spectrum are usually not complete enough for the accurate determination of glycan structure. When the algorithm proceeds, if the correct subtree structure is not included in candidate list in current round, the algorithm will not be able to correct such negative condition in the subsequent steps. Compared with CID spectrum, there are numerous quality peaks can be found in the larger m/z end of HCD spectrum, which provides more information about *Y*-ions of the glycopeptide. Besides, by controlling the fragmentation power, cleavages can happen at glycosidic bonds only while the attached peptide can be kept intact in HCD spectra. Since the mass of a *Y*-ion includes the mass of peptide, *Y*-ions and *B*-ions of the intact glycopeptide are separated in the two sides of a spectrum, which is favorable for spectra interpretation.

Based on the advantages of HCD fragmentation technique, which can provide quality spectra both in accuracy and resolution, we present a new method for glycan *de novo* sequencing from HCD spectra of N-linked glycopeptides. By utilizing the quality peaks in the larger m/z end of HCD spectrum, the glycan *de novo* sequencing problem is modelled as a top-down tree constructing process in a heuristic manner, in which the process starts at the glycosylation site and the peptide is regarded as the root of the tree. It will gradually construct the glycan tree from root to leaves until the mass of the glycan is reached.

In this chapter, we first present a mathematical model for the mass spectrometry based glycan *de novo* sequencing problem, and then propose a heuristic algorithm for glycan candidate generation. Next, we apply a re-evaluation scheme to re-rank the generated glycan candidates and report the top scored glycan that best fits the given mass spectrum. Moreover, the proposed algorithm is implemented to verify its effectiveness.

3.1 Notations and Problem Definition

A glycan structure was usually abstracted as a labelled rooted tree with node labels representing monosaccharide types [18, 20, 22]. In this section, the glycopeptide structure is represent as a tree and then a mathematical model for the glycan *de novo* sequencing problem is described.

3.1.1 Basic Notation

In an N-linked glycopeptide, usually only one glycan is attached to the peptide at the glycosylation site. The total mass value of a glycopeptide consists of three parts: residue mass of the peptide, residue mass of the glycan and an extra water. Let Σ_a be the alphabet of different types of amino acids and P be the peptide consisting of m amino acids. The string of the peptide can be represented as $P = a_1a_2 \dots a_m$. For an amino acid $a \in \Sigma_a$, we define $\|a\|$ as its residue mass value, then the residue mass of the peptide can be computed as $\|P\| = \sum_{1 \leq i \leq m} \|a_i\|$.

The sugar units that constitute a glycan through glycosidic bonds are called monosaccharides. It has been observed that some of the monosaccharides are epimers, which means that they differ only in their configurations rather than their mass. By *de novo* sequencing method, we cannot distinguish two epimers because they can hardly be separated from MS/MS spectra. Therefore, we only consider six types of common monosaccharides in this study based on the datasets we have. The name, abbreviation, composition, monoisotopic mass and symbol of these different types of monosaccharides are shown in Table 2.1. We use Σ_g to denote the alphabet of the different types (different mass) of monosaccharides. For a monosaccharide $g \in \Sigma_g$, $\|g\|$ is used to symbolize its residue mass value.

An N-linked glycan tree T is an unordered tree with its root linked to a peptide P . Each node of T represents a monosaccharide, labelled by an element from Σ_g . The degree of a

glycan tree is bounded by four because there are at most five linkages for one monosaccharide. Given a glycan tree T that include n monosaccharides, its mass can be represented as $\|T\| = \sum_{1 \leq i \leq n} \|g_i\|$. The actual mass of a glycopeptide G which consists of a glycan T and a peptide P is $\|G\| = \|P\| + \|T\| + \|H_2O\|$.

Assume that \mathcal{M} is used to denote the peak list of a glycopeptide spectrum, then we have $\mathcal{M} = \{(m_i, h_i) | i = 1, 2, \dots, n\}$. Each element (m_i, h_i) represents a peak in the spectrum, where m_i is the m/z value and h_i is the intensity of the peak. For a specific m/z value, if there is no peak in the spectrum \mathcal{M} , we consider that there exists a tuple $(m_i, h_i) \in \mathcal{M}$ with the intensity $h_i = 0$. The spectra need to be deisotoped and charge deconvoluted in a preprocessing step. The deisotoping is to convert signals from higher mass natural isotope peaks to their corresponding monoisotopic peak. The charge deconvolution will convert the multiply charged peaks to their singly charged equivalents [84]. In our experiment, data preprocessing is done by the software package PEAKS 6.0 [30]. Therefore, we assume that the spectrum \mathcal{M} only contains ions of charge one and the mass to charge ratio (m/z) of an ion is indeed equal to its mass value.

3.1.2 Mass Representation of Ion Fragments

In tandem mass spectrometry, glycopeptides will be fragmented and yield product ions and neutral fragments before they are transferred to the second mass analyzer. The fragmentations usually occur on the glycosidic bond of the glycopeptides. Theoretically, each fraction of the glycopeptide will generate one peak in the MS/MS spectrum. Six types of fragmented ions are commonly observed in the spectrum. From the reducing end, there are X-, Y-, and Z-ions; while in the non-reducing end, fragments are labelled with A-, B-, and C-ions. Figure 2.10 shows an example of a glycan and part of its fragmented ions. The A- and X-ions are generated from cross-ring cleavages. As we mentioned in the previous chapter, higher-energy collision dissociation (HCD) is currently a popular technique to generate glycopeptide MS/MS

for glycan structure identification. By controlling the fragmentation energy, the peptides of glycopeptides can be kept intact and we can obtain the tandem mass spectra in which a majority of peaks are generated from glycosidic bond cleavages. Practically, both *Y*-ions and *B*-ions dominate the fragment ions in HCD spectra with the former ion type generating more intensive peaks than the latter one, thus for simplicity we only consider *Y*-ions when generating possible glycan candidates. In the evaluation, we also take *B*-ions and internal fragment ions into consideration.

A glycan structure can be represented by a labelled rooted tree with node labels representing its monosaccharide types, as shown in Figure 3.1. Each monosaccharide of the glycan is abstracted as a node, and each glycosidic bond between two monosaccharide is represented as a linkage of the tree. In such representation, *B*-ions correspond to subtrees and *Y*-ions correspond to the remaining glycopeptide with the subtrees removed.

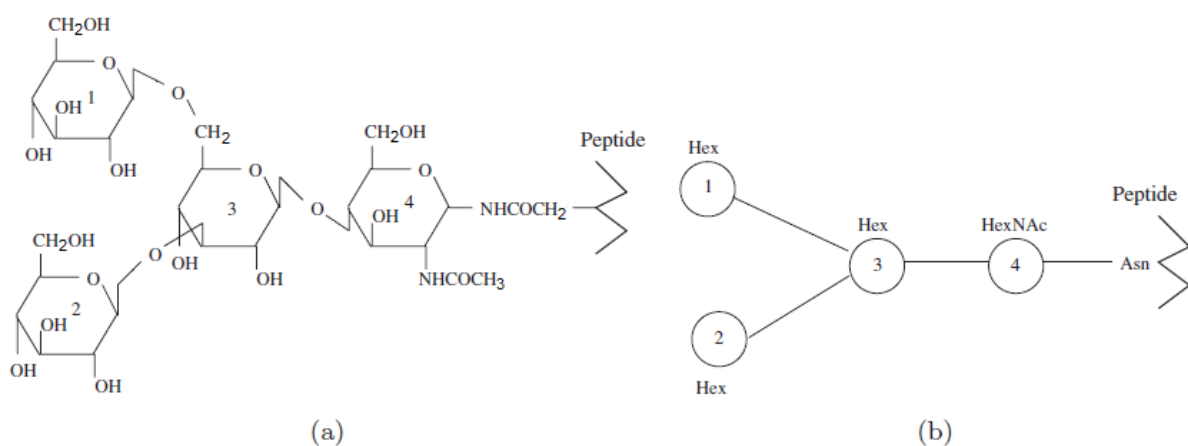


Figure 3.1: A glycopeptide and its tree representation. (a) A glycopeptide. The glycan attached to the peptide consists of four monosaccharides. (b) The abstract tree representation of the glycopeptide. Each monosaccharide is abstracted as a node in the tree representation. Each glycosidic bond between two monosaccharide is represented as a linkage of the tree.

Assume that t_i is the subtree of T rooted at i_{th} node and its mass is $\|t_i\|$, then the mass of the *B*-ion associated with t_i is $b_i = \|t_i\| + 1$.¹ A *Y*-ion corresponds to a subtree of T with the

¹There is a proton added to the ion in the ionization process.

peptide attached and we call it root-preserving subtree. B -ion and Y -ion are complementary to each other. Since a glycan structure is two-dimension, one Y -ion may complement with a set of B -ions. Let $y_{\{i_1, i_2, \dots, i_k\}}$ denote the mass value of a Y -ion corresponding to subtrees of T rooted at peptide P with $t_{i_1}, t_{i_2}, \dots, t_{i_k}$ removed, where $t_{i_1}, t_{i_2}, \dots, t_{i_k}$ are nonoverlapping subtrees of T respectively. We use y_0 to denote the mass of the Y -ion generated at the cleavage of glycosylation site, without glycan included. For a glycan tree structure constructed from its associated spectrum, its theoretical m/z values of singly charged Y -ions can be calculated by the enumeration of its root-preserving subtrees. The m/z value of a Y -ion is equal to the precursor mass value subtracting the mass of the removed monosaccharide residues. The m/z value of an internal fragment ion is equal to the total mass of the monosaccharide residues plus an additional proton. We use $I(G)$ to denote the set containing mass values of B -ions and Y -ions generated from a glycopeptide G with n monosaccharides, and it can be represented as follows,

$$I(G) = \bigcup_{1 \leq i \leq n} \{b_i\} \bigcup_{\{i_1, i_2, \dots, i_k\}} y_{\{i_1, i_2, \dots, i_k\}} \quad (3.1)$$

3.1.3 Problem Definition

The basic assumption of our model is that the greater number of high abundance peaks in the spectrum that are matched by those ions of a glycopeptide, the more likely the glycopeptide is the correct structure. More specifically, the evidence that a spectrum \mathcal{M} is generated from a glycopeptide G is that more and higher peaks in \mathcal{M} that are matched by theoretical ion fragments of G .

Because of the fact that the mass values acquired from mass spectrometers are not accurate, we use $\delta > 0$ to represent the maximum error bound of the mass spectrometers. Moreover, we use $M(G)$ to denote the set of peaks from the spectrum \mathcal{M} that match with the theoretical

ion mass values of G within the error tolerance δ ,

$$M(G) = \{(m_i, h_i) \in \mathcal{M} | \exists m \in I(G), |m - m_i| \leq \delta\} \quad (3.2)$$

For each peak in the set $M(G)$, a scoring function can be defined according to its mass value m and intensity h . Here we simply use $f(m, h)$ to denote the function and will discuss it later in detail.

Let $I(G)$ be all the possible ion mass values of a glycopeptide G , and $I(G)$ can be computed by Formula 3.1. $M(G)$ contains all the peaks in \mathcal{M} that can be explained by the ions of G . Intuitively, the more and higher peaks included in $M(G)$ indicates the more confident that \mathcal{M} is generated from G . Therefore, the GLYCAN DE NOVO SEQUENCING problem can be defined as follows: Given a spectrum \mathcal{M} , a precursor mass value M_p , a predefined error bound δ , and a peptide mass $\|P\|$, the objective is to construct a glycan tree T such that $|\|T\| + \|P\| + \|H_2O\| + 1 - M_p| \leq \delta$, and the score $S(T)$ is maximized,

$$S(T) = \sum_{(m_i, h_i) \in M(G)} f(m_i, h_i) \quad (3.3)$$

The summation above is used to evaluate how likely a glycopeptide matches with a spectrum. It is worth noticing that the peak list $M(G)$ is represented as a set meaning that different theoretical ion fragments with the same mass value matching the same peak in a spectrum will be counted only once in computing the scoring function.

3.1.4 Scoring Function

Several factors can be considered in the scoring function $f(m, h)$, such as mass value, intensity and ion types. Researchers can choose different factors to formulate the scoring function according to the instruments and configurations adopted in mass spectrometry experiments. For a peak with mass m and intensity h , a basic scoring function that only take the intensity of the

peak into consideration is defined as follows,

$$f(m, h) = \begin{cases} \log_2(h + \max(1 - h_{th}, 0)), & h \geq h_{th} \\ \log_2(0.5), & \text{otherwise} \end{cases} \quad (3.4)$$

The basic mechanism of the scoring function is that for each mass value m , it computes the reward/penalty that an ion has mass m . $h_{th} \geq 0.5$ is the threshold used to define whether a peak is useful or not. If there is a peak close to m with intensity $h \geq h_{th}$, the reward is equal to the logarithm of $h + \max(1 - h_{th}, 0)$. If the selected $h_{th} \geq 1$, the reward is equal to the logarithmic abundance of the peak directly. Otherwise, $1 - h_{th}$ will be added to h to guarantee the reward is always positive. If the intensity of a peak is less than h_{th} , we treat it to be a miss-match and a penalty score (a negative constant value) is imposed. In practice, we also consider the existence of N-linked glycan core structure when evaluating glycan structures. The peaks of each spectrum are normalized by dividing the intensity of highest peak and then scaled to the range of 0 ~ 100 before interpretation. In our experiment, h_{th} is set to 0.5, *i.e.*, if a peak with relative intensity less than 0.5% of the highest peak, there would be a penalty score $\log_2(0.5)$ assigned to that peak. The log function is to reduce the influence of those peaks with relatively high intensity.

3.2 Methods

3.2.1 Peptide Mass Inference

In the strategy that the intact glycopeptides are fragmented by tandem mass spectrometer, the mass values of peptides and their attached glycans are usually unknown prior to analysis. By controlling the fragmentation energy, HCD can break glycosidic bonds rather than peptide bonds, and keep the attached peptide intact. Thus we can infer peptide mass from HCD spectra generated from glycopeptides.

As mentioned above, N-linked glycans share a common core structure consisting of two N-Acetylglucosamine (GlcNAc) residues linked to a branched mannose (Man) triad, as shown in Figure 3.2. HCD usually favors the fragmentation of this core structure and generate corresponding Y -ions. In order to characterize the mass of peptide from spectrum, we need to find a series of peaks that can support this core structure (*i.e.*, peaks with mass values equal to peptide, peptide+GlcNAc, peptide+2GlcNAc, peptide+2GlcNAc+Man, peptide+2GlcNAc+2Man, peptide+2GlcNAc+3Man, respectively). More specifically, we need to identify a sequence of peaks in which the mass difference of two adjacent peaks equals to the mass of a certain monosaccharide residue.

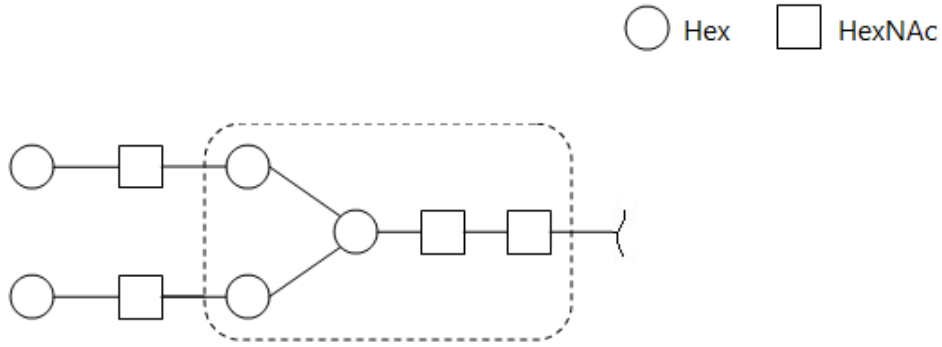


Figure 3.2: An example for core structure in an N-linked glycopeptide. The two GlcNAc linked to a branched Man triad in the dotted rectangle is the core structure.





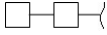
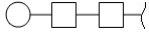
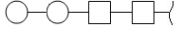
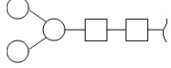
Let m_a, m_b be the mass values of a GlcNAc residue and a Man residue respectively. Suppose that the core structure of a glycopeptide is fragmented into a series of peaks in the spectrum \mathcal{M} at the m/z values in $M_{c_i} = \{m_i, m_i + m_a, m_i + 2m_a, m_i + 2m_a + m_b, m_i + 2m_a + 2m_b, m_i + 2m_a + 3m_b\}$. The mass representations for ion fragments of the core structure are summarized in Table 3.1. The peptide mass calculation problem can be defined as finding a peak in spectrum \mathcal{M} , such that the score of m_i calculated by the function $S(m_i)$ is maximized,

$$S(m_i) = \sum_{m \in M_{c_i}} f(m, h) \quad (3.5)$$

In Equation 3.5, the function $f(m, h)$ used to evaluate each peak in the set M_{c_i} is defined

in Equation 3.4. If there exists a peak that can support one of the fragments of core structure, a reward will be assigned to that peak. Otherwise, a penalty will be imposed. It is supposed that the mass value set M_{c_i} with larger score provides more evidence that there exists a core structure starting at m_i .

Table 3.1: Mass representations for the ion fragments of the N-linked glycan core structure

Fragment Composition	Mass Representation	Fragment Structure
GlcNAc	m_a	
Man	m_b	
Peptide	m_i	
Peptide+GlcNAc	$m_i + m_a$	
Peptide+2GlcNAc	$m_i + 2m_a$	
Peptide+2GlcNAc+Man	$m_i + 2m_a + m_b$	
Peptide+2GlcNAc+2Man	$m_i + 2m_a + 2m_b$	
Peptide+2GlcNAc+3Man	$m_i + 2m_a + 3m_b$	

In order to describe the algorithm more efficiently, we define a list of mass values $M_d = \{m_a, 2m_a, 2m_a + m_b, 2m_a + 2m_b, 2m_a + 3m_b\}$ to represent the mass differences between the peak at $m/z = m_i$ and other peaks in M_{c_i} . For each peak in the spectrum \mathcal{M} , we enumerate all its possible mass values associated with M_{c_i} by adding a mass value in M_d , and then calculate its score. After evaluation of all the peaks, the one with highest score will be reported and its mass is treated as the peptide residue mass. The pseudocode for this algorithm is shown in Algorithm 1.

3.2.2 Algorithm for Candidate Generation

It has been proved that the complexity of the glycan *de novo* sequencing problem is NP-hard, under the condition that each peak in spectrum cannot be repeatedly used when calculating

Algorithm 1 Peptide Mass Inference

INPUT: Given an MS/MS spectrum \mathcal{M} , a set of mass values M_d , a threshold h_{th} , and a predefined error bound δ .

OUTPUT: A peak $(m, h) \in \mathcal{M}$ with maximum score S_m .

```
1: for each  $(m_i, h_i) \in \mathcal{M} | h_i > h_{th}$  do
2:    $S_{m_i} = f(m_i, h_i)$ 
3:   for each  $\Delta m \in M_d$  do
4:     if  $\exists (m_j, h_j) \in \mathcal{M}$  s.t.  $|m_j - (m_i + \Delta m)| \leq \delta$  then
5:        $S_{m_i} += f(m_j, h_j)$ 
6: return  $(m, h)$  that has  $\max(S_m)$ 
```

scoring function [20]. Previous methods in [18] and [20] both constructed good solutions for smaller size trees and then assemble the reported trees into larger ones. Such strategy is suitable for glycan sequencing from CID spectra. While in HCD spectra, numerous quality peaks in the higher mass end can be used to explore new tree construction strategies. In this research, we provide a heuristic algorithm which construct the glycan tree from root to leaves based on HCD spectra.

A glycan tree with n vertices can be represented as $T = \langle v_1, v_2, \dots, v_n \rangle$, where v_i denotes a node of the tree T . If the whole peptide is treated as a node, the notation v_0 can represent such node as a single root. $d(v_i)$ is used to denote the degree of the subtree rooted at v_i , and $m(T)$ represents the summation of monosaccharide residue mass values of the glycan tree T . We use $F(n)$ to denote a set of glycan trees with n nodes,

$$F(n) = \{ T \mid T \text{ is a glycan tree, } |V_T| = n \} \quad (3.6)$$

Given a glycan tree T and a monosaccharide $g \in \Sigma_g$, let v be a node of T with less than four children, i.e., $d(v) < 4$. We use $T_{v \otimes g}$ to represent a new tree generated from T , where g is a monosaccharide added as a new node to the tree through node v . Thus, v becomes the parent node of g , and g becomes a leaf node of the newly constructed tree. In addition, we use

$T \otimes g$ to represent the set containing all of the possible glycan trees generated by adding g to

$$T = \langle v_1, v_2, \dots, v_n \rangle,$$

$$T \otimes g = \{T_{v_i \otimes g} \mid v_i \in V_T, d(v_i) < 4\} \quad (3.7)$$

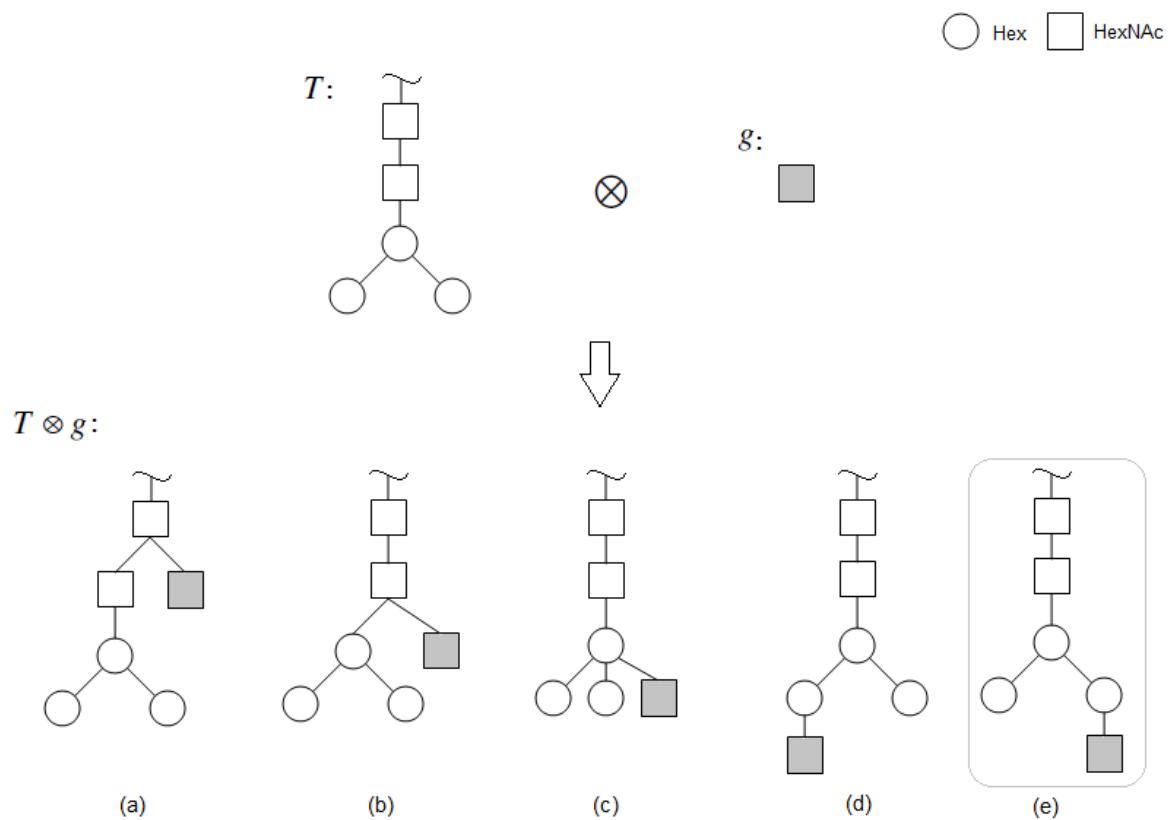


Figure 3.3: An example for new possible trees generated by adding a node g to a glycan tree T . The added node is shaded in grey.

Figure 3.3 shows an example for new possible trees generated by adding a node g to a glycan tree T . The original tree T contains five nodes, and the degree of each node is less than four. Thus the added node g can be attached to any of the nodes in T and five new trees can be generated theoretically, as shown in Figure 3.3 (a)-(e). However, considering that the tree representation we applied for a glycan structure is an unordered rooted tree, the two trees shown in Figure 3.3 (d) and (e) are actually the same. Therefore, there are four new trees generated. From this figure we know that it is necessary to eliminate the duplications of trees when generating new trees in order to control the size of candidate set. The algorithm to solve

this problem will be proposed in next section.

As we mentioned before, using HCD fragmentation method, the peptide can be kept intact during fragmentation. Thus, each Y -ion fragment corresponds to a subtree of the glycan tree T that rooted at v_0 . During the construction of the glycan tree, we need to find out all those subtrees to calculate the theoretical mass values of Y -ions. We use r to denote a root-preserving subtree of T and the mass value of its corresponding Y -ion is $\|r\| + \|P\| + \|H_2O\| + 1$.

Let $RPST$ denote the set of all the root-preserving subtrees of a glycan tree T , then we have,

$$RPST(T) = \{r \mid r \text{ is a root-preserving subtree of } T\} \quad (3.8)$$

Therefore, the set of root-preserving subtrees of $T \otimes g$ can be represented as follows,

$$RPST(T \otimes g) = \bigcup_{v_i \in V_T} RPST(T_{v_i \otimes g}) \quad (3.9)$$

For those root-preserving subtrees of a glycan tree T which include a node $v \in V_T$, we also define subtrees set $RPST(T, v)$ as follows,

$$RPST(T, v) = \{r \mid r \in RPST(T), v \in V_T\} \quad (3.10)$$

Specifically, $RPST(T, v)$ denotes a set of such root-preserving subtrees of a glycan tree T that include a path from root to node v . Figure 3.4 shows an example which demonstrates the set of root-preserving subtrees for the given glycan tree T . $RPST(T)$ contains five different root-preserving subtrees of T , as shown in Figure 3.4 (a)-(e). In the set $RPST(T)$, (c), (d), and (e) are the subtrees that contain the node v , thus they are in the set of $RPST(T, v)$.

Lemma 3.1 *Given a glycan tree T , and a monosaccharide $g \in \Sigma_g$, the set of root-preserving subtrees of the newly generated trees in $T \otimes g$ can be decomposed in the following way,*

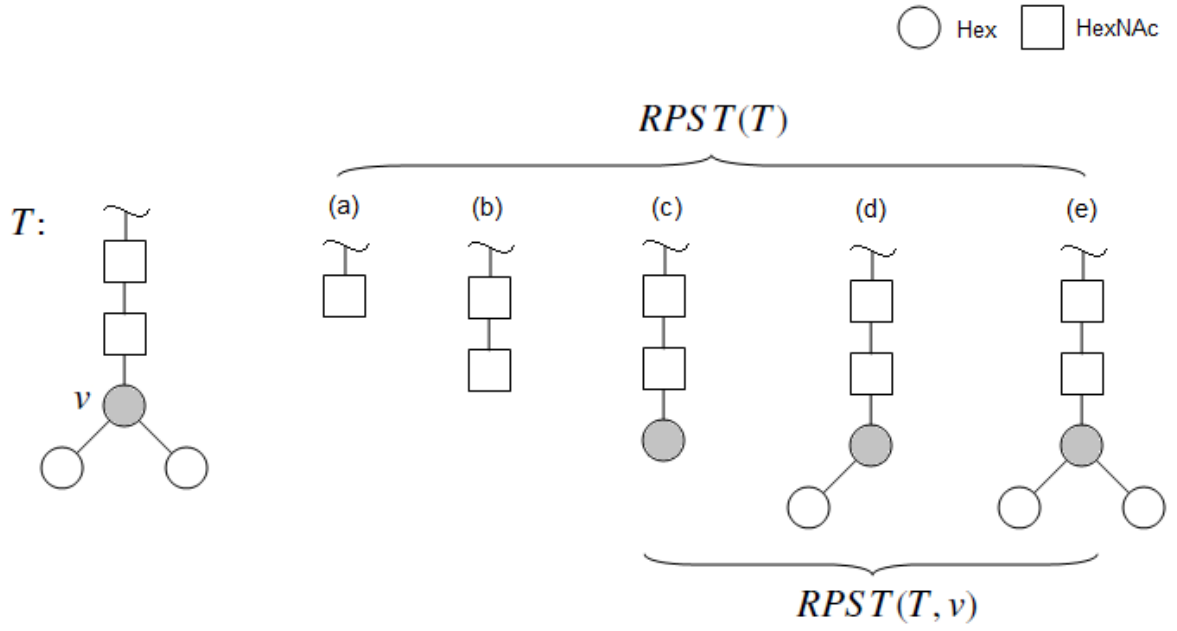


Figure 3.4: An example demonstrating the sets of $RPST(T)$ and $RPST(T, v)$ for the given glycan tree structure T . The node v in the tree T is shaded grey. There are five different root-preserving subtrees of T , in which three of them contain the node v and include a path from root to the node v .

$$RPST(T \otimes g) = RPST(T) \cup \bigcup_{\substack{v \in V_T, \\ r \in RPST(T, v)}} \{r_{v \otimes g} \mid d(v) < 4, v \in V_r\}$$

Proof First of all, let us consider a new tree $T_{v \otimes g}$ generated from a tree T and a monosaccharide g via node v . The set of root-preserving subtrees for $T_{v \otimes g}$ should contain all the subtrees from $RPST(T)$, because the original tree T is a subtree of $T_{v \otimes g}$. Since g is a new node added to the tree via node v , the subtrees in the set $RPST(T)$ do not contain g . Those newly generated root-preserving subtrees that contain g must also contain v , and they can be represented by the union $\bigcup_{r \in RPST(T, v)} \{r_{v \otimes g} \mid d(v) < 4\}$.

Thus we have the set of root-preserving subtrees of $T_{v \otimes g}$ as follows,

$$RPST(T_{v \otimes g}) = RPST(T) \cup \bigcup_{r \in RPST(T, v)} \{r_{v \otimes g} \mid d(v) < 4\} \quad (3.11)$$

Furthermore, the set $RPST(T \otimes g)$ corresponds to the calculation of $RPST(T_{v \otimes g})$ over all the nodes in V_T .

Based on the mathematical model described above, we now introduce our heuristic algorithm for the GLYCAN DE NOVO SEQUENCING problem. The glycan tree structure is gradually constructed by adding one node during each iteration in the computation. For an iteration with n nodes, a fixed number of glycan trees with highest scores are maintained in $F(n)$. M_{min} is set to be the minimum tree mass value in $F(n)$. Under the condition that even if we add a monosaccharide with the smallest mass value to M_{min} , the result is greater than $M' = M_p - m_p - 19$, then the program will be terminated. $F(n)$ is computed in the following two steps:

1. We compute a set of candidate glycan structures based on the equation

$$F_c(n) = \bigcup_{g \in \Sigma_g} \bigcup_{T \in F(n-1)} T \otimes g;$$

2. Compute the score of each structure in $F_c(n)$ by evaluating how its theoretical ion masses match with peaks in mass spectrum \mathcal{M} according to the scoring function defined in equation ??, then put the top $|F|$ glycans in $F(n)$ and remove those glycans with low scores.

During the construction process, if the mass of one generated glycan satisfies the desired glycan mass value M' , then this glycan will be put into the candidate results set R . When the program finished, a fix number of glycans sorted by scores from high to low can be obtained in R . Algorithm 2 presents the pseudocode for the algorithm of glycan candidates generation described above.

Algorithm 2 Glycan Candidates Generation by *De Novo* Sequencing

INPUT: Given a spectrum \mathcal{M} , and glycopeptide precursor mass value M_p , and peptide mass value m_p , and a predefined error bound δ .

OUTPUT: A set R consists of candidate glycan structures with their scores, and each glycan tree T in R satisfies $||T|| + ||P|| + ||H_2O|| + 1 - M_p \leq \delta$.

```
1:  $M' = M_p - m_p - 19$ ,  $M_{min} = m_p$ 
2: while  $M_{min} + \min||g|| \leq M'$  do
3:    $F_c(n) = \emptyset$ 
4:   for  $T \in F(n-1)$  do
5:     for  $v_i \in V_T$  do
6:       for  $g \in \Sigma_g$  do
7:          $T' = T_{v_i \otimes g}$ 
8:          $RPST(T') = RPST(T)$ 
9:         for  $r \in RPST(T)$  do
10:          if  $v_i \in V_r$  and  $d(v_i) < 4$  then
11:             $RPST(T') = RPST(T') \cup r_{v_i \otimes g}$ 
12:           $F_c(n) = F_c(n) \cup T'$ 
13:   Score each glycan tree in  $F_c(n)$  according to  $RPST(T)$ , put top  $|F|$  in  $F(n)$ 
14:   Set  $M_{min}$  to be the minimum mass value of trees in  $F(n)$ 
15:   Select the trees from  $F(n)$  that satisfies mass requirement and put them in  $R$ 
```

3.2.3 Algorithm for Elimination of Duplicate Trees

The glycan structures are abstracted as unordered rooted trees in our method. The direct computation of the set $T \otimes g$ and $RPST(T \otimes g)$ from Equation 3.7 and Lemma 3.1 may generate duplicate or isomorphic trees. Two trees are said to be isomorphic if that one tree can be mapped into the other by permuting the order of the children of vertices [85]. We use a candidate set with fixed size to include all the trees generated in each iteration. The removal of identical trees should be taken into consideration. Otherwise, the size limit of the candidate

set maintained during computation would be reached quickly, yet the correct result may not be included. To solve this problem, an algorithm similar to the tree isomorphism determination algorithm described in [85] has been designed to eliminate the duplications of trees in a set.

For each newly generated tree, we first compare its tree mass and tree size with the trees in the set. If two trees have the same tree mass and size, we need to assign a string to each of them according to its structure and node information. Then we compare the two strings to determine whether they are isomorphic. If they are isomorphic, we do not need to add the new tree into the set. The string for a glycan tree T is obtained as follows.

1. Assign to each leaf of T a string of integer representing its node type. The correspondence between the integers and node types can be found in Figure 3.5.
2. Inductively, assume that all vertices of T at height $i + 1$ have been assigned strings.
3. Assign to each nonleaf of T at height i a string by the following way: for each vertex v at height i take the integer i_0 assigned to v according to the correspondence shown in Figure 3.5 to be the first component of the string, and then sort all the strings associated with its children in alphabetic order. The alphabet consists of integers assigned to different types of monosaccharides as well as the left and right parentheses. The sorted strings are then combined together and form a tuple. On completion of this step, each nonleaf v of T at height i will have a string $i_0(s_1)(s_2)...(s_k)$ associated with it, where $s_1, s_2, ..., s_k$ are the strings, in nondecreasing order, associated with the children of v .
4. Repeat step 3 until the root of T is assigned. That string can be used to represent the tree T and used to compare with other trees.

In order to facilitate the understanding of the algorithm, Figure 3.5 illustrates the assignment of strings to the vertices of one tree. As can be seen from the figure, the string assignment

is iteratively from leaf to root. At last, the string assigned to the root node of the tree is treated as the string of the tree. The strings generated according to the procedure described above for two isomorphic trees are identical, because the strings assigned to each child of one node in the same level are sorted and concatenated together. By comparing the strings associated with two glycan tree structures, we can determine whether these two trees are isomorphic or not.

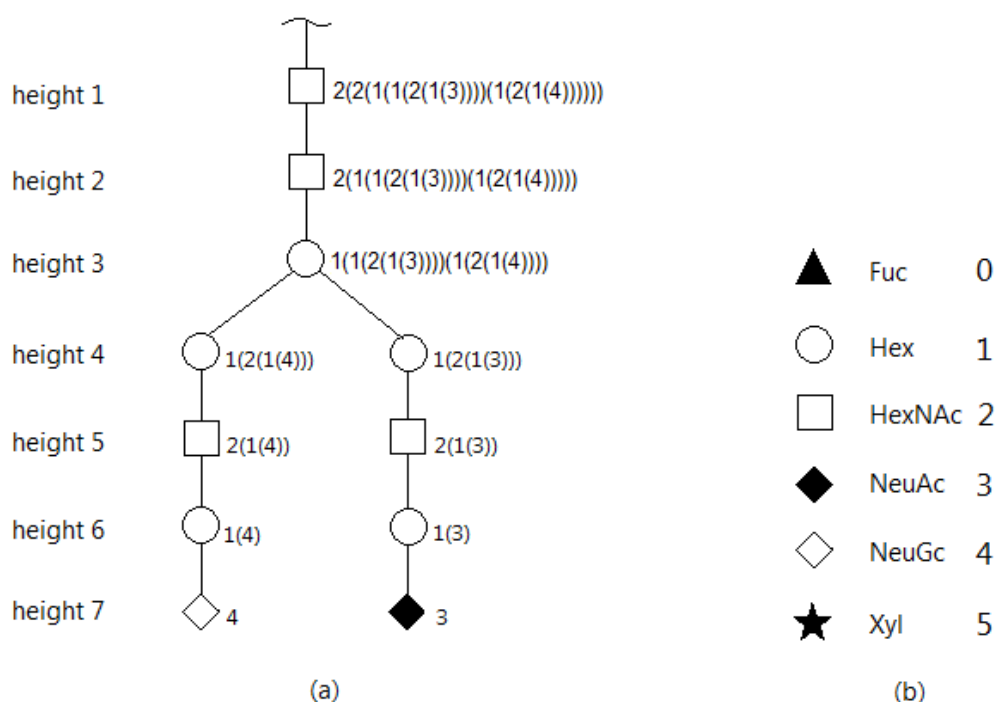


Figure 3.5: (a) Strings assigned by isomorphic tree elimination algorithm. The final string associated with the tree is $2(2(1(2(1(3))))(1(2(1(4))))))$. (b) Integers assigned to different types of monosaccharides.

The detailed procedure of two tree isomorphism determination is summarized in Algorithms 3 and 4. The following notions are used in the algorithms. T is denoted as a glycan tree with height H and mass $m(T)$. $|T|$ is the size of tree T and S_T is the string assigned to T . For a node v in the tree T , $C(v)$ represents the children set of v .

Algorithm 3 String Representation for A Glycan Tree

INPUT: Given a glycan tree T , tree height H , and the integers assigned to different types of monosaccharides.

OUTPUT: A string representing T according to its structure and node type.

```
1:  $h \leftarrow H$ 
2: while  $h > 0$  do
3:   for each node  $v$  in  $V_i$  do
4:     Set the integer  $i_0$  assigned to its type as its string  $s_v$ 
5:     if  $v$  has children  $|C(v)| > 0$  then
6:       Sort the strings assigned to each child in lexicographic order, denoted as
          $(s_1)(s_2)\dots(s_k)$ 
7:       Set  $s_v$  by concatenating  $i_0$  and  $(s_1)(s_2)\dots(s_k)$  together
8:      $h --$ 
9: return the string assigned to the root node of  $T$ 
```

3.2.4 Complexity Analysis

The string assignment of an n -vertex labelled tree and determination of the isomorphism of two trees can be done in $O(n)$ time [85]. Therefore, we can eliminate the duplications of trees in a set in $O(nN \log N)$ time, where N is the size of the set. We give the complexity of the proposed Algorithm 2 and Algorithm 4 in the following Theorem 3.1.

Theorem 3.1 *Algorithm 2 and Algorithm 4 compute glycan candidates from tandem mass spectra by de novo sequencing in time bounded by*

$$O(n^4 \times c \times |\Sigma_g| \times |F| \times \log(|F|))$$

Proof During each round of the tree construction when adding a new node to each tree of the previous candidate list, there will be $n \times |\Sigma_g|$ possible newly generated trees. If the size of each

Algorithm 4 Tree Isomorphism Determination

INPUT: Two glycan tree T_1 and T_2 , and an error range δ_m .

OUTPUT: TRUE if T_1 and T_2 are identical; otherwise FALSE.

```
1: if  $|m(T_1) - m(T_2)| \leq \delta_m$  then
2:   Return FALSE
3: if  $|T_1| \neq |T_2|$  then
4:   Return FALSE
5: if  $S_{T_1} == \text{null}$  or  $S_{T_2} == \text{null}$  then
6:   Invoke Algorithm 3 to set the string for  $T_1$  or  $T_2$ 
7: if  $S_{T_1} \neq S_{T_2}$  then
8:   Return FALSE
9: else
10:  Return TRUE
```

candidate list is denoted as $|F|$, then there will be $n \times |\Sigma_g| \times |F|$ new trees generated. The time complexity for eliminating duplications of these new trees is bounded by $O(n^2 \times |\Sigma_g| \times |F| \times \log(|F|))$.

For all the newly generated trees, each one is scored according to the matching between its root-preserving subtrees against the peaks in spectrum. The number of root-preserving subtrees of a specific tree is exponential to its size theoretically, which is bounded by $O(2^n)$. However, in practice during our computation for generating glycan structures, the maximal size of *RPS T* was less than 50. The main reason is that the degree of most nodes in a glycan tree is less than four and the size of each glycan tree is small. For simplicity, we consider that the time complexity for scoring each candidate tree structure is bounded by $O(cn)$, where c is a constant. In total, there will be maximum of n rounds of adding new nodes. Overall, the time complexity of the algorithm proposed above is $O(n^4 \times c \times |\Sigma_g| \times |F| \times \log(|F|))$, where n is the total number of the vertices for a glycan tree, which is less than 20 practically.

3.2.5 Re-evaluation Scheme

During the previous step of glycan candidates generation, the ions we consider in the scoring function are mainly Y -ions. Because of the complexity of glycan tree structures, two situations may exist in the higher scored candidates. One is that many similar glycan structures are reported by the algorithm with highest score, and the true structure is among them but can not be distinguished. The other situation is that the target glycan is ranked in higher scored place but not the first. Consider these situations, a more rigorous re-evaluation scheme should be applied to filter candidate glycan structures and reduce the number of candidates.

In the post-processing section, each candidate glycan structure is theoretically fragmented at glycosidic bonds and ion types we take into consideration here are B -ions and Y -ions as well as internal fragment ions. The internal fragment ions refer to those fragments generated from more than one cleavage on a glycan structure simultaneously. In addition, a new scoring function is defined to re-evaluate glycan candidates according to their ion peaks. Let (m_B, h_B) be the mass value and intensity for a B -ion. And (m_Y, h_Y) , (m_I, h_I) are for a Y -ion and an internal ion, respectively. The new score for a glycan tree candidate $S_{re}(T)$ is calculated as follows,

$$S_{re}(T) = \alpha \sum f(m_{Bi}, h_{Bi}) + \beta \sum f(m_{Yj}, h_{Yj}) + \theta \sum f(m_{Ij}, h_{Ij}) \quad (3.12)$$

In Equation 3.12, the function $f(m, h)$ used to evaluate a peak is the same as the one defined in Equation 3.4. The parameters α, β, θ are used to adjust the relative weight of different ion types. HCD spectra are featured with the predominance of Y -ions, while B -ions and internal ions can be occasionally observed in the low m/z end of the spectra. Thus, if a peak in the experimental spectrum supports a Y -ion of a glycan candidate, that peak will be assigned more weight than that supports a B -ion or an internal ion. In our experiment, α, β , and θ are empirically set to 0.5, 1.0, and 0.3, respectively.

After re-evaluation with a more stringent scheme, the glycan candidate structures are re-ranked according to their new scores. Those glycan structures with higher re-ranking scores are selected as the final candidates.

3.3 Experiments and Results

The approach proposed above was implemented in the experiments to test its performance and the top 1000 candidates were selected during each iteration of computation, *i.e.*, $|F| = 1000$. The error bound $\delta = 0.2$ Da were used in the experiment.

3.3.1 Datasets

The glycopeptide samples used in the experiments were derived from three kinds of protein samples: Alpha-1-acid glycoprotein of *Bos taurus* (Bovine), Ovomucoid of *Gallus gallus* (Chicken), and Ig gamma-3 chain C region of *Homo sapiens* (Human). Experiments were carried on a Thermo Scientific Orbitrap Elite hybrid mass spectrometer and HCD fragmentation technique was used.

The newly developed software tool GlycoMaster DB [16] was used for comparison. GlycoMaster DB can analyze mass spectra produced with HCD fragmentation and identify N-linked glycans by searching against the glycan structure database GlycomeDB [72]. The main reason we choose a database searching method for comparison is that the algorithms mentioned in [18, 20–22] which designed based on *de novo* sequencing method cannot handle glycopeptide data or can only analyze CID spectra. Besides, the results identified by database searching method are relatively reliable.

Our experimental dataset contains 46 HCD spectra of glycopeptides, which were identi-

Table 3.2: Performance of our algorithm compared with GlycoMasterDB

Rank ¹	Number of glycans	Percentage(%)
1	35	76.09
2	6	13.04
3 ~ 10	1	2.17
11 ~ 20	1	2.17
>20	1	2.17
Can't find	2	4.35

fied by GlycoMaster DB from the collected 1168 MS/MS spectra. The reported glycan structures were used to benchmark the performance of our proposed method.

3.3.2 Experimental Results

For each MS/MS spectrum in the dataset, top 10 candidates of glycan structures were reported by GlycoMaster DB. Among those results, the highest ranked glycan structure was treated as the reference structure, and this structure was compared with all the results constructed by our algorithm. Table 3.2 shows the ranking status of the reference structures observed in our reported results for those 46 MS/MS spectra.

As one can see from Table 3.2, there are 35 glycans with highest scores generated by our proposed method have the same structures as those top-ranked glycans interpreted by GlycoMaster DB. Besides, there are 6 glycans identified from their associated spectra by GlycoMaster DB with highest scores ranked second place in our results. Therefore, if the case that the corresponding reference structure ranking top two in our reported results is deemed correct, then the accuracy rate of our proposed method can reach to 89.13%. Among the results that the reference structures ranked greater than 10, for one entry its reference structure ranked 13

¹"Rank" refers to the ranking status of the **reference structure** (the top glycan structure reported by GlycoMaster DB) in our result for a spectrum.

and for the other entry is 32, which is the lowest rank observed in our results. However, after observing that entry with lowest ranking, we can find that the two top ranked glycan structures reported by our method and GlycoMaster DB were quite similar. Besides, the second ranked glycan structures reported by both methods were identical. And the scores of those top two results reported by GlycoMaster DB were indeed very close, which are 25.97 and 25.81 respectively. In order to see the results intuitively, their structures are listed in Figure 3.6. In the figure, structure (a) is the one ranked the first place in the results of GlycoMaster DB, which is ranked 32 in our results. (b) shows the glycan structure ranked top by our method and (c) shows the one ranked second in the results of both methods.

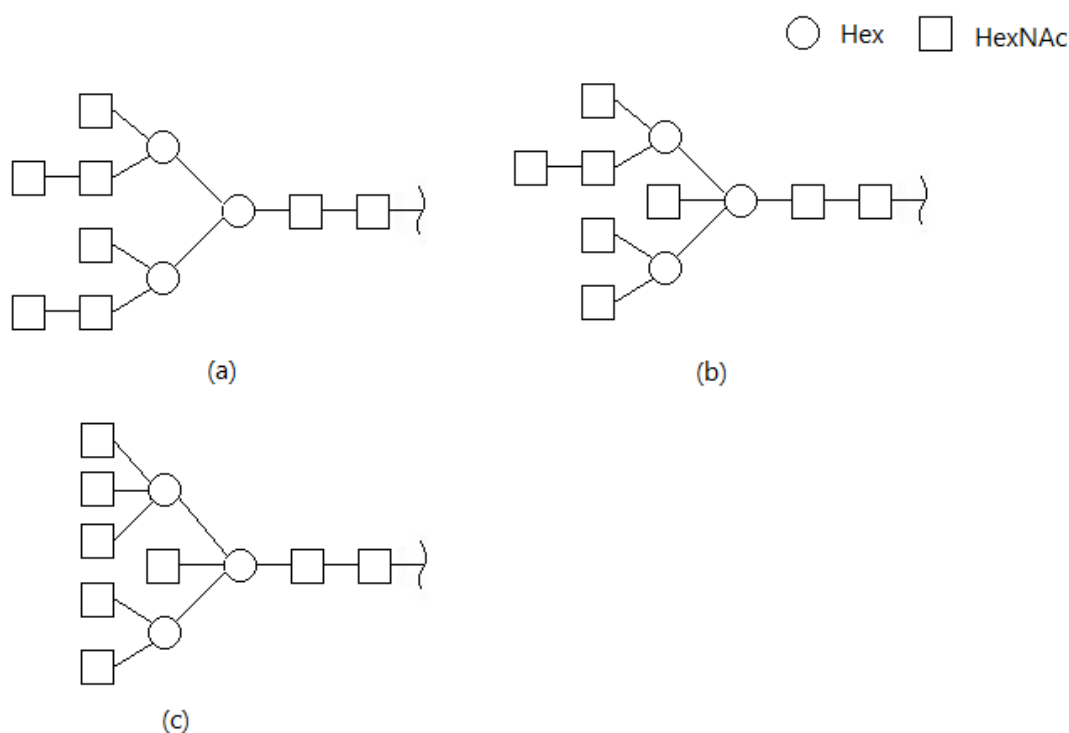


Figure 3.6: Comparison of the top two glycan structures reported by GlycoMaster DB and our method for a certain spectrum. (a) Ranked No.1 and No.32 in the results of GlycoMaster DB and our method respectively. (b) Top ranked result in our method. (c) Ranked second place in both methods.

There are two entries that the reference glycan structures cannot be observed in the results provided by our proposed algorithm. However, our reported glycans with highest score were only partially different from the related reference structures. Figure 3.7 shows the difference of

these two pairs of results.

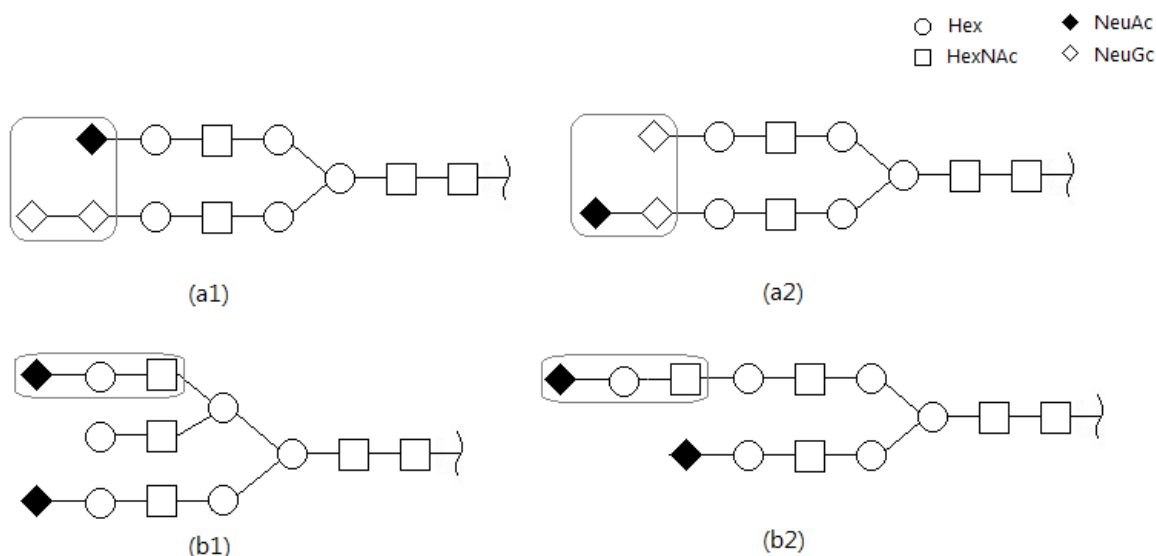


Figure 3.7: Comparison of two pairs of glycan structures identified by GlycoMaster DB and our method respectively. (a1) and (a2) are the top ranked glycans identified from the same spectrum and (b1) and (b2) are identified from another spectrum. Both (a1) and (b1) are the results reported by GlycoMaster DB, while (a2) and (b2) are reported by our method.

Each pair of glycans in the same row shown in Figure 3.7 were interpreted from the same HCD spectrum. The first glycan in each row was identified by GlycoMaster DB with highest scores, and the second one in each row was ranked first in the results of our method. As one can see from the figure, the two glycans in the same row have high resemblance and one can be converted to the other through a few steps of operations. This indicates that although our algorithm did not find out the same glycan structures as GlycoMaster DB did, our results are quite similar to the reference structures. After analyzing the two reference structures and those associated spectra, we find that the precursor mass of both glycans are relatively large, which are more than 3000 Da. However, in the experiment, fragment ions with mass values greater than 2000 Da can rarely be observed. Thus, it is difficult to construct the whole tree structure correctly using *de novo* sequencing method due to the lacking of ion information. Although we can take advantage of the *b*-ions from the smaller mass side of the spectrum, the information is still not adequate to infer the whole tree structure in some of the cases in practice. For

example, in the spectrum associated with glycan (a1) and (a2) shown in Figure 3.7, there are two peaks at m/z 292.10 and 308.10, which can be interpreted as monosaccharides NeuAc (◆) and NeuGc (◇), but there is no peak that can support the disaccharide NeuGc-NeuGc (◇-◇) in (a1). There are indeed two peaks at m/z 657.2365 and 673.2311, which can be interpreted as the trisaccharides HexNAc-Hex-NeuAc (□-○-◆) and HexNAc-Hex-NeuGc (□-○-◇) respectively. However, the intensity of the peak supporting HexNAc-Hex-NeuGc is 434.49655, which is much higher than the other one 192.08842. Thus, glycan structure in (a2) is more likely ranked higher than that in (a1) during the tree construction process. Accordingly, we believe that using more types of ions such as internal ions during construction may improve the results. In general, our reported results were reasonable because they were supported well by each corresponding spectrum.

3.4 An Improved Result

In our previous results, it is observed that the main difference between our top ranked glycan structures and the ones reported by GlycoMaster DB appear at the leaf end of the trees. As shown in Figure 3.7 (a1) and (a2), structure (a1) can be changed into (a2) if the two monosaccharide NeuAc and NeuGc change position with each other. By investigating the spectra, we can find that there are several peaks that support *B*-ions and internal ions. Therefore, we improve our previous method by taking *B*-ions and internal ions into consideration during the tree construction procedure.

Most spectra of N-linked glycopptides have peaks corresponding to oxonium ions formed by one to three monosaccharide residues in low m/z region. For each spectrum, we maintain a set which contains five possible linear trisaccharides well supported by significant peaks. During each round of generating new trees, we check to see if there is a leaf-containing linear trisaccharide structure of the current tree same as one in the set. A bonus will be added to this

tree in this case.

Table 3.3: Performance of improved method compared with previous *de novo* sequencing method

Rank ¹	De Novo Sequencing Method		Improved Method	
	Number	Percentage(%)	Number	Percentage(%)
1	35	76.09	40	86.96
2	6	13.04	1	2.17
3	0	0.00	1	2.17
4 ~ 10	1	2.17	1	2.17
> 10	2	4.35	2	4.35
Can't find	2	4.35	1	2.17

Better results were obtained based on the improvements, as shown in Table 3.3. As can be seen from the table, there are 40 top ranked glycans reported by our improved method having the same structures as those top-ranked glycans interpreted by GlycoMaster DB. Compared with the original *de novo* sequencing algorithm, the improved method has higher accuracy.

There are two entries that the reference glycan structures can not be observed in the results reported by our previous algorithm as shown in Figure 3.7. After improving the algorithm, the structure (a1) shown in the figure can be found in the results although it is not ranked very high. The reason is clear. According to the mechanism of the improvement, the trisaccharide HexNAc-Hex-NeuAc ($\square-\bigcirc-\blacklozenge$) shown in (a1) can help the structure to earn a bonus during the construction, which increases the possibility of the structure (a1) to be included in the candidate list.

3.5 Conclusion and Discussion

In this chapter, a mathematical model for glycan *de novo* sequencing problem is represented as a tree construction problem under the condition that the score of the tree is maximized

¹“Rank” refers to the ranking status of the **reference structure** (the top glycan structure reported by GlycoMaster DB) in our results for a spectrum.

according to its associated spectrum. We propose a heuristic algorithm for glycan candidates generation from HCD MS/MS spectra. Each tree is constructed from root (peptide) to leaves in a heuristic manner by utilizing the quality peaks in the higher mass end of the spectra. Furthermore, in order to filter candidate glycans more effectively and reduce the number of candidates, a re-evaluation scheme is applied and those glycans with higher re-ranking scores are selected as final candidates.

The proposed approach is applied to identify glycan structures from 46 HCD tandem mass spectra of N-linked glycopeptides. The performance of the proposed approach is compared with a successful database searching method, GlycoMaster DB. For each spectrum, the top ranked glycan structure reported by GlycoMaster DB is treated as the reference structure and its ranking status in our results is considered as criteria for assessment. Experimental results have shown that the results of our proposed method are comparable with GlycoMaster DB, with the accuracy rate reaching to 89.13%. In terms of those reference structures which are not ranked higher or not found in our results, our top reported glycan structures are quite similar to them. It indicates that our results can be applied in some database searching method by assisting the filtration of tentative candidates.

Comparing with database searching method, *de novo* sequencing method require high-quality MS/MS spectra for a reliable analysis. It is still a challenging problem in glycan *de novo* sequencing. In next chapter, we will combine our *de novo* sequencing method with database searching method to improve the accuracy of glycan structure characterization from MS/MS spectra.

Chapter 4

Glycan Structure Identification using Database Search and *De Novo* Sequencing

Currently, two commonly used computational approaches for glycan structure identification with tandem mass spectrometry are database search and *de novo* sequencing. To some extent, *de novo* sequencing method is more difficult than database search method, since it does not rely on any glycan database. Glycan identification by *de novo* sequencing can only rely on the information obtained from tandem mass spectra, without the assistance of any structural knowledge of glycans. Generally, database search method has the ability to obtain more reliable results than *de novo* sequencing method. However, a distinctive advantage of glycan *de novo* sequencing is that it can identify some new glycan structures which are not included in the glycan database. With the development of mass spectrometry technology, more quality mass spectra will be produced, which can be used for glycan *de novo* sequencing to get more accurate results. Both of these two approaches have advantages and disadvantages, thus a better way is to combine them together for interpretation of glycan structures from tandem mass spectra.

In this chapter, we proposed a new method for matching the input spectra with glycan structures acquired from glycan database by incorporating a *de novo* sequencing assisted rank-

ing scheme. The *de novo* sequencing results used to screen the candidates from database are acquired by the method proposed in the previous chapter. The new approach typically interprets the spectral data in three steps: (1) peptide mass inference, (2) glycan candidates selection from glycan database and raw score calculation, (3) filtration of glycan candidates by incorporating *de novo* sequencing results. In addition, we improved the efficiency of this *de novo* sequencing assisted database search method by combining the glycan candidates acquired from both methods and re-evaluating them to get better results. Experimental results showed that our proposed methods not only can identify glycans from tandem mass spectra with high accuracy, but also has the potential for finding new glycan structures that not included in the current glycan database.

4.1 Preliminaries

4.1.1 Notations and Problem Formulation

In an N-linked glycopeptide, a glycan is attached to a peptide at the glycosylation site. The total mass of a glycopeptide consists of the residue mass of peptide, glycan and an extra water. The glycan structure can be abstracted as a labelled rooted unordered tree, where each node label represents a monosaccharide type [18, 20, 22]. The degree of such glycan tree is bounded by four because one monosaccharide can have five linkages at most. Similar to the notations in Chapter 3, six types of common monosaccharides are considered as shown in Table 2.1. Epimers, which differ only in their conformation rather than their mass, are not taken into consideration, because they can hardly be distinguished in a MS/MS spectrum. Assume that in an N-linked glycopeptide G , a glycan tree T is an unordered tree with its root linked to a peptide P and the tree includes n monosaccharides. We use Σ_g to denote the alphabet of different types (mass) of monosaccharide. For a monosaccharide $g \in \Sigma_g$, $\|g\|$ is used to symbolize its residue mass value. The mass of tree T can be represented as $\|T\| = \sum_{1 \leq i \leq n} \|g_i\|$, and the mass of the

glycopeptide G is $\|G\| = \|P\| + \|T\| + \|H_2O\|$.

In MS/MS, cleavages of glycosidic bonds result in B/Y -ions, C/Z -ions, and A/X -ions. Figure 2.10 shows an example of a glycopeptide and its ion fragments. The A/X -ions are generated by cross-ring cleavages under a relatively high collision energy [61]. In HCD spectra, Y -ions dominate all the fragment ions. Research also shows that peaks representing internal fragment ions can be observed in the low m/z region [86]. The internal fragment ions refer to those fragments generated by more than one cleavages on a glycan structure simultaneously. Therefore, we take B/Y -ions as well as internal ions into consideration when calculating the theoretical fragment ions of a glycan structure in database. In the tree representation, B -ions correspond to subtrees and Y -ions correspond to the remaining glycopeptide with one or more subtrees removed, while internal ions are part of a subtree. Here, we use $I(G)$ to denote the set of mass values of B/Y -ions and internal ions generated from a glycopeptide G . For the glycans in database, the Y -ion mass values of their corresponding glycopeptides can be computed by adding the mass of attached peptides to their own Y -ions.

Assume that a spectrum \mathcal{M} is generated by the fragmentation of a glycopeptide G , and \mathcal{M} is represented by a peak list $\mathcal{M} = \{(m_i, h_i) | i = 1, 2, \dots, n\}$, where m_i and h_i represent the mass and the intensity of a peak in the spectrum, respectively. In addition, we use M_p to denote the precursor mass value of the glycopeptide. It can be obtained from the m/z and charge state z reported by the experiment instrument. For a specific m/z value, if there is no corresponding peak in spectrum \mathcal{M} , we consider there to be a tuple $(m_i, h_i) \in \mathcal{M}$ with its intensity $h_i = 0$. Intuitively, the more and higher peaks in the spectrum \mathcal{M} can match with ion fragments of the glycopeptide G , the more likely that \mathcal{M} is generated from G . More formally, $M(G)$ is adopted to denote the set of peaks from spectrum \mathcal{M} that match with the theoretical ion mass values of G within an error tolerance $\delta > 0$, then we have

$$M(G) = \{(m_i, h_i) \in \mathcal{M} | \exists m \in I(G), |m - m_i| \leq \delta\} \quad (4.1)$$

The problem of glycan structure identification from a spectrum using database search method is to find the best matching glycan by comparing the MS/MS spectrum against all glycan structures from a glycan database. Thus, the GLYCAN DATABASE SEARCHING PROBLEM can be formulated as follows: given a MS/MS spectrum \mathcal{M} , a precursor mass value M_p , a predefined error bound δ , a peptide mass $\|P\|$, and a glycan database D , the objective is to find a glycan structure T from D that maximize the value of a specific scoring function $S(\mathcal{M}, T)$, such that $\|T\| + \|P\| + \|H_2O\| + 1 - M_p \leq \delta$. The scoring function $S(\mathcal{M}, T)$ evaluates how likely the glycopeptide associated with the glycan tree T matches with the spectrum \mathcal{M} . The scoring function $S(\mathcal{M}, T)$ can be defined as follows,

$$S(\mathcal{M}, T) = \sum_{(m_i, h_i) \in M(G)} f(m_i, h_i) \quad (4.2)$$

In Equation 4.2, the function $f(m, h)$ is used to calculate the score for a fragment ion matched by a peak with mass value m and intensity h in spectrum \mathcal{M} . We will discuss the definition of this function in next section. It is worth noticing that different theoretical ion fragments with the same mass value that match the same peak in a spectrum will be counted only once during the computation of the scoring function.

4.1.2 Scoring Function

The scoring function is essential for best glycan candidate selection that we should pay special attention to. Usually the intensity value of the matched peaks, the fragment ion types, as well as the mass errors are taken into consideration when deriving a reasonable function $f(m, h)$ to measure the reward or penalty each matched peak can impose on the overall matching. The function defined in this method is based on the one proposed in Chapter 3. We improved Equation 3.4 by incorporating the consideration of mass error [30]. The score for a fragment ion matched by a peak with mass value m and intensity value h is calculated as follows,

$$f(m, h) = \begin{cases} \log_2(h + \max(1 - h_{th}, 0)) \times \exp(-(\frac{m-m'}{\delta})^2), & h \geq h_{th} \\ \log_2(0.5), & \text{otherwise} \end{cases} \quad (4.3)$$

In Equation 4.3, m' is the mass value of an ion, m is the mass value of the observed peak for that ion, and δ is the mass error tolerance of the mass spectrometer. The exponential factor in the equation reflects the mass error. h denotes the intensity value of the observed peak. $h_{th} \geq 0.5$ is the threshold set to define whether the peak is useful or not. If the intensity of the peak is less than h_{th} , it is treated as a miss-match and a penalty score is imposed. In practice, the peaks of each spectrum are normalized by dividing the intensity of highest peak and scaled to range of 0 ~ 100 before interpretation. In our experiment, h_{th} is set to 0.5, *i.e.*, if a peak with relative intensity less than 0.5%, there would be a penalty score $\log_2(0.5)$ assigned to that peak.

4.2 Methods

The glycan identification by *de novo* sequencing assisted database search method mainly contains three steps. The first step is inferring the mass value of peptide from the spectrum. Since we are trying to identify the glycan structure from the spectrum of intact glycopeptide, the mass value of the peptide in the glycopeptide may be unknown in advance. We need to infer the mass value of peptide first, and then calculate the mass value of the attached glycan. This step has been described in Chapter 3, thus omitted in this chapter. The second and third steps are glycan candidates selection from database and filtration of candidates with the assistance of *de novo* sequencing results respectively. In the following of this section, these two steps will be illustrated.

4.2.1 Glycan Candidates Selection and Raw Score Calculation

In this step, firstly the glycan candidates which satisfy the mass value requirement will be selected from the database. Secondly, the raw score of each candidate will be calculated by matching its theoretical fragment ions with the given spectrum.

As mentioned in the section of *Notations*, the total mass of a glycopeptide consists of a peptide mass, a glycan mass, and a water. The theoretical mass value of the peptide can be inferred from the spectrum based on the method described in Chapter 3 by taking advantage of core structure in an N-linked glycopeptide. Then we can obtain glycan mass by subtracting the inferred peptide mass and an $\|H_2O\|$ from the input molecular mass. Moreover, since our research focuses on N-linked glycosylation, we extracted all the N-linked glycans from the glycan database GlycomeDB in advance. Given an inferred peptide mass value m'_p , a predefined mass tolerance Δ , and a precursor mass value M_p , those N-linked glycans whose mass values m_T satisfy the equation $|M_p - m'_p - \|H_2O\| - 1 - m_T| \leq \Delta$ will be selected from the database as the tentative candidates.

The correctness of a glycan structure is justified based on the intuition that it should produce more and higher-intensity peaks and generate a higher score than other structure does. In order to evaluate to what extent a candidate glycan matches with the spectrum, its simulated theoretical MS/MS is used to compare against experimental MS/MS and the raw matching score $S_{raw}(T)$ is given based on Equation 4.3. $S_{raw}(T)$ denotes the summation of all theoretic fragment ion scores for a glycan tree T . The ion types we take into consideration in this study are B -ions, Y -ions as well as internal fragment ions. Let (m_B, h_B) be the mass value and intensity for a B -ion. Meanwhile, we use (m_Y, h_Y) and (m_I, h_I) to represent a Y -ion and an internal ion respectively. The raw score for a glycan tree candidate $S_{raw}(T)$ is calculated as follows,

$$S_{raw}(T) = \alpha \sum f(m_{Bi}, h_{Bi}) + \beta \sum f(m_{Yj}, h_{Yj}) + \theta \sum f(m_{Ik}, h_{Ik}) \quad (4.4)$$

In Equation 4.4, the function $f(m, h)$ used for each peak is the same as the one defined in Equation 4.3. The parameters α, β, θ are used to adjust the relative weight of different ion types. In our experiment, α, β , and θ are empirically configured to be 0.5, 1.0, and 0.25 respectively.

The raw score serves for the purpose of pre-evaluation of each candidate glycan structure in database. However, only using raw score to rank all the candidates is not sufficient, because it could happen that several similar glycan structures match to the same group of peaks in the spectrum that we cannot decide which one is the best matching structure. Therefore, a further filtration strategy is needed to obtain better results.

4.2.2 Algorithm for Filtration by *De Novo* Sequencing

In Chapter 3, we formulated the problem of glycan *de novo* sequencing from MS/MS spectra, and proposed a heuristic algorithm for identifying glycan from HCD MS/MS spectra of N-linked glycopeptides. The algorithm has the ability to construct the best matching glycan tree structures from the given glycopeptide MS/MS spectra. Although not all the correct glycans can be identified as the best matching structures for the spectra used in the experiment, our top reported glycan structures are quite similar to those correct ones. In this step, we will employ the *de novo* sequencing results to screen the candidates acquired from glycan database.

In the *de novo* sequencing algorithm, the top 1000 ranked glycan structures will be reported for each spectrum, among which those glycans with score greater than half of the highest score are selected for the database candidates filtration. We use $L_{nov} = \{R_1, R_2, \dots, R_m\}$ to denote such a sorted list of glycans selected from *de novo* sequencing result, where $R_i (1 \leq i \leq m)$ represents a glycan structure with score larger than half of the score of top ranked glycan. The rank of R_i in the list is denoted as $rank(R_i)$. Besides, we use $L_{db} = \{Q_1, Q_2, \dots, Q_n\}$ to represent the list of glycan candidates selected from the glycan database. Then the filtration strategy

proceeds as follows.

1. For each glycan Q_j in L_{db} , compare it with a glycan R_i in L_{novo} . For a pair (Q_j, R_i) , we calculate the similarity score $S_{sim}(Q_j, R_i)$, based on which we then calculate the comparison score $S_{comp}(Q_j, R_i)$ according to the following equation,

$$S_{comp}(Q_j, R_i) = S_{sim}(Q_j, R_i) \times \exp\left(\frac{1}{rank(R_i)}\right) \quad (4.5)$$

2. Repeat step 1 until the comparison score $S_{comp}(Q_j, R_i)$ for each glycan structure R_i in list L_{novo} against Q_j is calculated. Then sort R_i in L_{novo} according to their comparison score and select top K glycans to form a new list $L'_{novo}(j)$.
3. The score of a glycan Q_j in database list L_{db} is readjusted based on two parts. One is its raw score calculated by matching its theoretic mass spectrum with experimental spectrum, the other is the comparison score calculated according to its similarity to the glycans R_k in the list $L'_{novo}(j)$. The equation for calculating the score of Q_j is shown as follows,

$$S(Q_j) = \log(S_{raw}(Q_j) \times \frac{1}{K} \times \sum_{k=1}^K S_{comp}(Q_j, R_k)) \quad (4.6)$$

4. Repeat above steps until each Q_j in the database list has been scored according to Equation 4.6. Then we sort these glycans in decreasing order and select the top F glycans as the final filtration result.

Algorithm 5 summarizes the strategy of glycan candidates filtration by incorporating the results from our *de novo* glycan sequencing method. The following notations are used in the algorithm. S_{raw} denotes the raw score of a glycan structure from database calculated according to Equation 4.4. S_{sim} is the similarity score between two trees from database and *de novo* sequencing results respectively. In next section we will discuss how to calculate this similarity score. S_{comp} is the comparison score between two trees which considers not only the similarity between them but also the rank of the tree in *de novo* sequencing result.

Algorithm 5 Filtration of glycan candidates

INPUT: Given an MS/MS spectrum \mathcal{M} , a sorted list of glycans $L_{novo} = \{R_1, R_2, \dots, R_m\}$ from *de novo* sequencing, a list of glycan candidates $L_{db} = \{Q_1, Q_2, \dots, Q_n\}$ from database.

OUTPUT: A ranked list of glycans with size F from database list L_{db} .

- 1: **for** each j from 1 to n **do**
 - 2: calculate its raw score $S_{raw}(Q_j)$
 - 3: **for** each i from 1 to m **do**
 - 4: calculate the similarity score $S_{sim}(Q_j, R_i)$
 - 5: $S_{comp}(Q_j, R_i) = S_{sim}(Q_j, R_i) \times \exp(\frac{1}{rank(R_i)})$
 - 6: sort all R_i in L_{novo} and select top K scored glycans
 - 7: $S(Q_j) = \log(S_{raw}(Q_j) \times \frac{1}{K} \times \sum_{k=1}^K S_{comp}(Q_j, R_k))$
 - 8: sort all Q_j in L_{db} in decreasing order and select top F as filtration result.
-

4.2.3 Algorithm for Calculating Similarity Between Labelled Unordered Trees

In the filtration strategy, the similarity score S_{sim} between two glycan structures is calculated based on the comparison of two glycans from the database list and the *de novo* list respectively. As mentioned in the section of *Notations*, a glycan is usually abstracted as a rooted labelled unordered tree with degree less than or equal to four. In order to calculate similarity score, we first compute the editing distance between two labelled unordered glycan trees, and then convert the distance measure to a similarity measure.

Labelled unordered trees are rooted trees whose nodes are labelled and in which only ancestor relationships are significant while the left-to-right order among siblings is not significant. That is, the children of a node in a labelled unordered tree do not have an ordering [87]. There are three kinds of operations for unordered trees: (1) changing a node n means changing the label on n ; (2) deleting a node n means making the children of n become the children of the parent of n and then removing n ; (3) inserting n as a child of n' will make n the parent of some

subset of the current children of n' .

Suppose we have a numbering for each unordered glycan tree. Let $t[i]$ be the i th node of tree T in the given numbering. $T[i]$ be the subtree rooted at $t[i]$ and $F[i]$ be the unordered forest by deleting $t[i]$ from $T[i]$. Let θ denote the empty tree and λ denote the null node. Let T_1, T_2 be two trees, a is either λ or a label of a node in T_1 and b is either λ or a label of a node in T_2 . Three kinds of edit operations are considered, which are change ($a \rightarrow b$), delete ($a \rightarrow \lambda$), and insert ($\lambda \rightarrow b$). Let γ be a cost function that assigns to each edit operation $a \rightarrow b$ a non-negative real number $\gamma(a \rightarrow b)$. We constrain γ to be a distance metric. That is,

1. $\gamma(a \rightarrow b) \geq 0; \gamma(a \rightarrow a) = 0$
2. $\gamma(a \rightarrow b) = \gamma(b \rightarrow a)$
3. $\gamma(a \rightarrow c) \leq \gamma(a \rightarrow b) + \gamma(b \rightarrow c)$

Let S be a finite sequence s_1, s_2, \dots, s_k of edit operations. Then the editing distance between two labelled unordered trees T_1 and T_2 is defined as follows,

$$d(T_1, T_2) = \min_S \{\gamma(S) | S \text{ is a finite sequence of edit operations transforming } T_1 \text{ to } T_2\}$$

In addition, a triple (M_e, T_1, T_2) is defined to be an editing distance mapping from T_1 to T_2 , where $M_e \subset [1 \dots |T_1|] \times [1 \dots |T_2|]$ is any set of pairs of integers (i, j) satisfying: For any pairs (i_1, j_1) and (i_2, j_2) in M_e ,

1. $i_1 = i_2$ if and only if $j_1 = j_2$ (one-to-one);
2. $t_1[i_1]$ is an ancestor of $t_1[i_2]$ if and only if $t_2[j_1]$ is an ancestor of $t_2[j_2]$ (ancestor order preserved).

Given S , a sequence s_1, s_2, \dots, s_k of edit operations from T_1 to T_2 , there exists a mapping M_e from T_1 to T_2 such that $\gamma(M_e) \leq \gamma(S)$ [87]. The relation between the editing distance and the editing distance mapping is

$$d(T_1, T_2) = \min_{M_e} \{\gamma(M_e) | M_e \text{ is an editing distance mapping from } T_1 \text{ to } T_2\}$$

In [87], Zhang et al. showed that the computation of the editing distance between unordered labelled trees is NP-complete, even if the trees are binary trees with a label alphabet of size two. Motivated by this, a new editing based distance between unordered trees is defined in [88]. The intuitive idea of the new editing distance is based on a restriction of the mapping between two trees, which is two separate subtrees of tree T_1 should be mapped to two separate subtrees in tree T_2 . Suppose a triple (M, T_1, T_2) is a restricted editing distance mapping from T_1 to T_2 , then any set of pairs of integers (i, j) should satisfy the following conditions. First, for any pairs (i_1, j_1) and (i_2, j_2) in M , they satisfy the restrictions required in editing distance mapping. Second, for any triples (i_1, j_1) , (i_2, j_2) and (i_3, j_3) in M , $\text{lca}(t_1[i_1], t_1[i_2])$ is a proper ancestor of $t_1[i_3]$ if and only if $\text{lca}(t_2[j_1], t_2[j_2])$ is a proper ancestor of $t_2[j_3]$, where lca represents least common ancestor.

Based on the restricted mapping, the new editing based distance between T_1 and T_2 is defined as follows,

$$D(T_1, T_2) = \min_M \{\gamma(M) | M \text{ is a restricted editing distance mapping transforming } T_1 \text{ to } T_2\}$$

The algorithm we used to compute the editing distance between labelled unordered trees T_1 and T_2 is similar to the algorithm proposed in [89]. The recursion functions used in the algorithm are shown in Equation 4.7 and Equation 4.8.

$$D(F_1[i], F_2[j]) = \min \begin{cases} D(F_1[i], \theta) + \min_{1 \leq s \leq n_i} \{D(F_1[i_s], F_2[j]) - D(F_1[i_s], \theta)\}, \\ D(\theta, F_2[j]) + \min_{1 \leq t \leq n_j} \{D(F_1[i], F_2[j_t]) - D(\theta, F_2[j_t])\}, \\ \min_{MM(i,j)} \gamma(MM(i, j)). \end{cases} \quad (4.7)$$

$$D(T_1[i], T_2[j]) = \min \begin{cases} D(T_1[i], \theta) + \min_{1 \leq s \leq n_i} \{D(T_1[i_s], T_2[j]) - D(T_1[i_s], \theta)\}, \\ D(\theta, T_2[j]) + \min_{1 \leq t \leq n_j} \{D(T_1[i], T_2[j_t]) - D(\theta, T_2[j_t])\}, \\ D(F_1[i], F_2[j]) + \gamma(t_1[i] \rightarrow t_2[j]). \end{cases} \quad (4.8)$$

In Equation 4.7, $MM(i, j)$ is defined as a maximum matching on a restricted mapping between $F_1[i]$ and $F_2[j]$. In [89], the computation of $\min_{MM(i,j)} \gamma(MM(i, j))$ was defined to be the minimum cost maximum bipartite matching problem and reduced to the minimum cost maximum flow problem by adding two empty trees to $F_1[i]$ and $F_2[j]$ respectively. However, in our glycan tree comparison problem, we use a more straightforward way to calculate this value. We know in prior that the degree of a glycan tree is bounded by 4, thus we can simply try every permutation between the children of $t_1[i]$ and the children of $t_2[j]$ to find the minimum cost for this case. Let n_i and n_j be the number of children of $t_1[i]$ and $t_2[j]$ respectively and suppose $n_i \geq n_j$. There are totally $P(n_i, n_j)$ permutations to be considered. One exemplary permutation between $F_1[i]$ and $F_2[j]$ can be seen in Figure 4.1. Compared with the original algorithm of constructing a minimum cost maximum flow network, the simple idea of trying every permutation is much easier to implement.

The editing based distance between two trees is a distance metric. To simplify our calculation, we then convert the distance metric into a similarity metric. The transformation proposed in [90] was used. Let $D(T_1, T_2)$ be the editing distance between tree T_1 and T_2 , and θ is an empty tree, then the similarity score $S_{sim}(T_1, T_2)$ between T_1 and T_2 can be calculated as follows,

$$S_{sim}(T_1, T_2) = \frac{D(T_1, \theta) + D(T_2, \theta) - D(T_1, T_2)}{2} \quad (4.9)$$

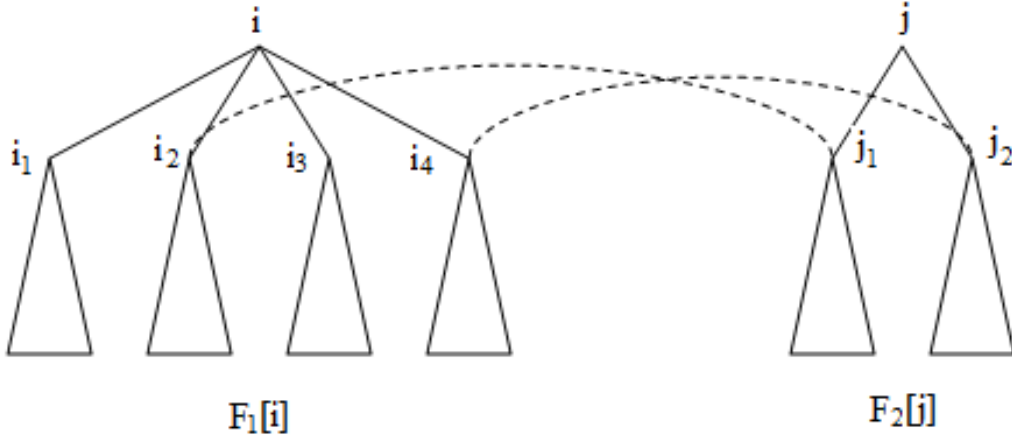


Figure 4.1: An example of permutation between $F_1[i]$ and $F_2[j]$, where the two subtrees $T_2[j_1]$ and $T_2[j_2]$ in $F_2[j]$ are mapped to the two subtrees in $F_1[i]$ which can reach the minimum cost, say $T_1[i_2]$ and $T_1[i_4]$ respectively, and all the other subtrees in $F_1[i]$ are deleted.

4.2.4 Complexity Analysis

In Algorithm 5 that computes the editing based distance, the complexity of computing $D(T_1[i], T_2[j])$ is bounded by $O(n_i + n_j)$. In our case, n_i and n_j both are less or equal to 4, therefore the total times of calculation is not larger than 8. The complexity of computing $D(F_1[i], F_2[j])$ is bounded by $O(n_i + n_j)$ plus the complexity of trying every permutation in case 3 of Equation 4.7, whose value is bounded by $24 = P(4, 4)$. Hence for any pair i and j , the complexity of computing $D(T_1[i], T_2[j])$ and $D(F_1[i], F_2[j])$ is bounded by $40 = 8 + 8 + 24$, and this can be denoted as a constant $C (C \leq 40)$. Therefore the complexity of our algorithm for computing editing based distance is $O(C \times |T_1| \times |T_2|)$. To sum up, the time complexity for the filtration in total is $O(n \times m \times C \times |T_1| \times |T_2|)$, where n is the size of glycan candidates from glycan database, and m denotes the size of the glycan list from *de novo* sequencing.

4.3 Experiments and Results

Based on the proposed approach, we have developed a software program, GlycoNovoDB, for glycan structure identification from HCD glycopeptide spectra. For each spectrum, the top 20 scored glycans were reported as the final candidates, *i.e.*, $F = 20$.

4.3.1 Datasets

The glycopeptide samples used in the experiments were derived from three kinds of protein samples: Alpha-1-acid glycoprotein of *Bos taurus* (Bovine), Ovomucoid of *Gallus gallus* (Chicken), and Ig gamma-3 chain C region of *Homo sapiens* (Human). Experiments were carried on a Thermo Scientific Orbitrap Elite hybrid mass spectrometer and HCD fragmentation technique was used.

The software tool GlycoMaster DB [16] and our previously proposed *de novo* sequencing algorithm [24] were used for comparison. GlycoMaster DB can analyze mass spectra produced with HCD fragmentation and identify N-linked glycans by searching against the glycan structure database GlycomeDB [72]. Our previous method doesn't use any database knowledge, it can also identify glycans from spectra effectively with an accuracy rate of 89.13% if top two ranked glycans are deemed correct. The performance of these two methods as well as our new developed program will be tested on the same dataset.

Our experimental dataset contains 46 HCD spectra of glycopeptides, which were identified by GlycoMaster DB from the collected 1168 MS/MS spectra. The reported glycan structures were used to benchmark the performance of our proposed methods.

Table 4.1: Performance of *de novo* sequencing method and GlycoNovoDB compared with GlycoMaster DB

Rank ¹	<i>De Novo</i> Sequencing Method		GlycoNovoDB	
	Number	Percentage(%)	Number	Percentage(%)
1	40	86.96	45	97.83
2	1	2.17	1	2.17
3	1	2.17	0	0.00
4 ~ 10	1	2.17	0	0.00
> 10	2	4.35	0	0.00
can't find	1	2.17	0	0.00

4.3.2 Experimental Results

For each MS/MS spectrum in the dataset, GlycoMaster DB reported 10 glycan candidates with highest score. Among these results, the top ranked glycan structure was treated as the reference structure in our experiment. The reference structure was compared with all the results reported by our *de novo* sequencing algorithm as well as our new approach GlycoNovoDB. Table 4.1 shows the ranking status of reference structures observed in our reported results for those 46 MS/MS spectra.

From Table 4.1, we can see that there are 40 glycans with highest scores generated by our previous *de novo* sequencing method have the same structures as those top-ranked glycans interpreted by GlycoMaster DB. While compared with our new method, GlycoNovoDB, 45 glycans ranked at the first place reported by GlycoMaster DB are also ranked top by our new method. In our previous *de novo* sequencing method, there is one entry that the reference glycan structure cannot be observed in the reported result. However, all the reference structures for each given spectrum can be identified by GlycoNovoDB and ranked at least top two. In general, it is easy to see that our new proposed approach, GlycoNovoDB, can provide results with much higher accuracy than the previous *de novo* sequencing method. The reasons are obvious. One

¹"Rank" refers to the ranking status of the **reference structure** (the top glycan structure reported by GlycoMaster DB) in our results for a spectrum.

is that the glycan database can provide relatively reliable glycan structure information that can improve the accuracy of the identification results. Another reason may be that although for several spectra, their reference glycan structures identified by GlycoMaster DB are not ranked the first place in our *de novo* sequencing results, the glycans constructed by our previous method with higher score are only partially different from the related reference structures [24]. In this case, with the support of our *de novo* sequencing results, the glycan candidates selected from the database can be filtered efficiently and a better result can be reached.

For the 46 HCD spectra of glycopeptides, there are 6 spectra that GlycoMaster DB reported more than one top-ranked glycan structures with the same score, which means there are several different glycan structures sharing the same highest score that cannot be distinguished. However, this did not happen in our program GlycoNovoDB. By interpreting each mass spectrum manually, there indeed exists some evidence that GlycoNovoDB can report more confident results than GlycoMaster DB. One example can be seen in Figure 4.2. The two glycan structures shown in the figure are interpreted from the same spectrum and both ranked top with the same score in GlycoMaster DB result. While in our result, glycan (a) has higher score than glycan (b) and is ranked the first place. By investigating the ion fragments of these two glycans and comparing them with the peaks in their associated spectrum, we have found that there is one peak at m/z 569.221 with intensity 2192.233, which can be interpreted as an ion consists of residues of two HexNAc (\square) and a Hex (\circ). From the two glycans shown in Figure 4.2, it can be seen that glycan (a) is easy to generate a b -ion corresponding to that peak, as shown in the dotted rectangle. However, in order to produce the same peak, glycan (b) needs three cleavages to generate an internal ion, which is much more difficult. Therefore, we believe that glycan (a) is more likely to be the correct glycan structure for the given spectrum.

method and database search method by merging the two sets of candidates together and designing more efficient scoring function to identify glycans that are in the database with high accuracy or provide new glycan structures that are not in database confidently.

4.4.1 Methods

In the *de novo* sequencing method proposed in Chapter 3, each glycan candidate is constructed from root to leaves in a heuristic manner by utilizing the quality peaks in the higher mass end of the HCD MS/MS. On one hand, the construction process does not rely on any knowledge of glycan database, thus the glycan candidates generated from *de novo* sequencing method may contain novel glycan structures. On the other hand, the glycan candidates selected from glycan database by GlycoNovoDB have high accuracy with the assistance of structure information when the correct structure is indeed included in the database. Our main idea is to combine the two sets of candidates acquired from *de novo* sequencing method and database search method together, and then evaluate the score of each candidate using the same scoring function to find out the glycan structure that is best matched with the given spectrum.

In order to evaluate all the glycan candidates fairly and effectively, a strict scoring function should be designed. The fragment ion types we considered here are *Y*-ions, *B*-ions, and internal ions. The score function used to calculate the score for a fragment ion matched by a peak is the same as Equation 4.3. Then the scoring function to evaluate to what extent a candidate glycan structure matches with the given spectrum is generally based on the summation of all theoretic fragment ion scores of the glycan structure. We also consider the percentage of matched peaks among all the peaks in the spectrum. Equation 4.10 calculates the score of a glycan structure T matched with a spectrum.

$$S_{imp}(T) = (\alpha \sum f(m_{Bi}, h_{Bi}) + \beta \sum f(m_{Yj}, h_{Yj}) + \theta \sum f(m_{Ik}, h_{Ik})) \times \frac{N_{match}}{N_{all}} \quad (4.10)$$

In Equation 4.10, the function $f(m, h)$ is the same as the one defined in Equation 4.3. (m_B, h_B) , (m_Y, h_Y) , and (m_I, h_I) represent B -ion, Y -ion, and internal fragment ion respectively. N_{match} denotes the number of peaks that matched with ions, and N_{all} is the total number of the peaks in the spectrum. The parameters α, β, θ are used to adjust the relative weight of different ion types, since a certain fragment technique usually does not produce the same amount of different fragment ions. Here, α, β , and θ are empirically set to be 0.5, 1.0, and 0.25 respectively.

Given a spectrum \mathcal{M} , a glycan database D , the improved method for glycan identification proceeds as follows,

1. Use the *de novo* sequencing algorithm to generate a list of candidates, and select top 20 ranked glycan into the sorted list C_{nov} .
2. Use the database search method GlycoNovoDB to generate the glycan candidates from database D . Similarly, top 20 ranked glycans will be selected into the list C_{db} .
3. If the two top ranked glycans in C_{nov} and C_{db} have the same structure, then we treat this glycan structure as the one that is best matched with the given spectrum and output as the result.
4. Otherwise, the theoretical fragment ions of each glycan candidate in the list C_{nov} and C_{db} will be considered. By matching the fragment ions against the peaks in the spectrum \mathcal{M} , the score of each candidate can be calculated according to Equation 4.10.
5. Sort all the glycan candidates according to its score S_{imp} . If the top ranked glycan is from the list C_{db} , it is treated as the correct structure for the given spectrum. Otherwise,

we believe that the database may not contain the correct glycan structure associated with M , and our *de novo* sequencing results provide a good reference.

4.4.2 Experimental Results

The dataset we used to test the performance of this improved method is the same as the one used in the experiment for testing GlycoNovoDB. The dataset contains 46 HCD spectra of glycopeptides, which were identified by GlycoMaster DB from the collected 1168 MS/MS spectra.

Among all these 46 spectra of glycopeptides, there are 40 spectra that *de novo* sequencing algorithm and GlycoNovoDB reported the same top-ranked glycan structure for each spectrum. Thus we believe that these 40 spectra are confidently interpreted. The glycan structure identified for each spectrum is treated as the correct glycan that is associated with the given spectrum.

For the left 6 spectra in the dataset, there are two spectra for which the glycans identified by GlycoNovoDB have higher scores than those reported by the *de novo* sequencing algorithm in the new scoring system. Figure 4.3 and 4.4 show the glycan structures reported by GlycoNovoDB and *de novo* sequencing method for each of the two spectra respectively. As shown in Figure 4.3, the two glycan structures are interpreted from the same spectrum. Glycan (a) is identified by GlycoNovoDB while glycan (b) is identified by the *de novo* sequencing algorithm, and glycan (a) has higher score than (b) after calculation based on the same scoring function. By checking their associated mass spectrum manually, we believe that this result is meaningful. In the spectrum, there is one peak at m/z 569.2213 with intensity 2844.5134, which can be interpreted as an ion consists of residues of two HexNAc(\square) and a Hex (\bigcirc). From the two glycan structures shown in Figure 4.3, it can be seen that glycan (a) is more likely to generate a

B-ion corresponding to that peak, which is circled in the dotted rectangle. There exists a peak at m/z 772.30096 in the spectrum, which can be considered as a support for the ion shown in the dotted rectangle in glycan (b). However, the intensity of that peak is only 51.14829. After normalization by dividing the intensity of highest peak (117376.984), it is scaled to 0.043576, which is less than the threshold $h_{th} = 0.5$. Thus there would be a penalty score assigned to that peak, which may reduce the score of the glycan structure (b).

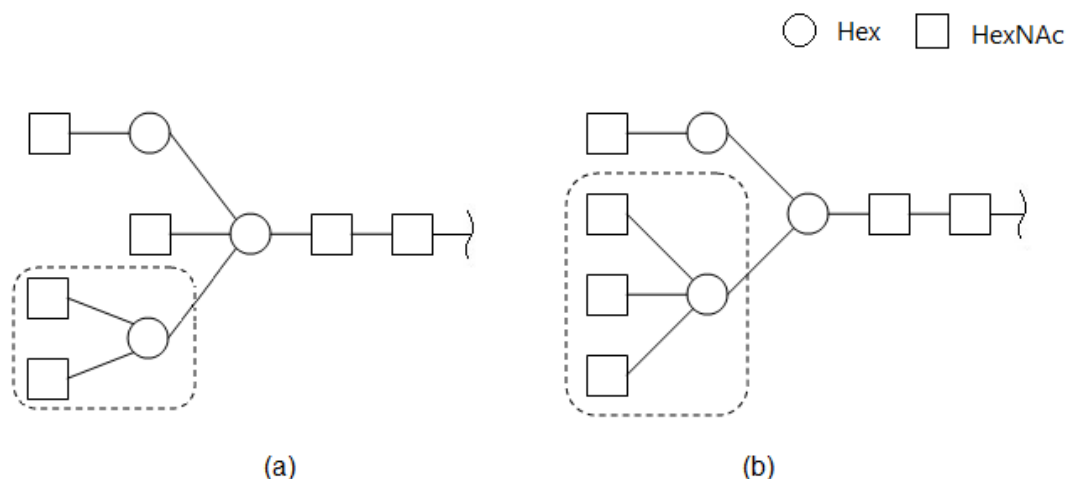


Figure 4.3: Comparison of the two glycan structures identified from the same spectrum by GlycoNovoDB and *de novo* sequencing algorithm respectively. Glycan (a) identified by GlycoNovoDB has higher score than glycan (b) which is reported by *de novo* sequencing method.

Similar to Figure 4.3, Figure 4.4 shows another example that the glycan structure identified by GlycoNovoDB has higher score than the one identified from the same spectrum by the *de novo* sequencing method. By investigating the theoretical ion fragments of the two glycans shown in Figure 4.4 and comparing them with the peaks in their corresponding spectrum, it can be found that there is a peak at m/z 325.1133 with intensity 145.17714. This peak can be interpreted as an ion consists of residues of two Hex (○), which is more likely to be a *B*-ion as shown in the dotted rectangle in glycan (a). While for the glycan (b) shown in the figure, it will require at least two cleavages to generate such ion fragment. Besides, glycan (b) is quite possible to generate a *B*-ion which has three residues of Hex (○), but we can not find any peak can support this fragment ion, since there is no peak at m/z 487.1662.

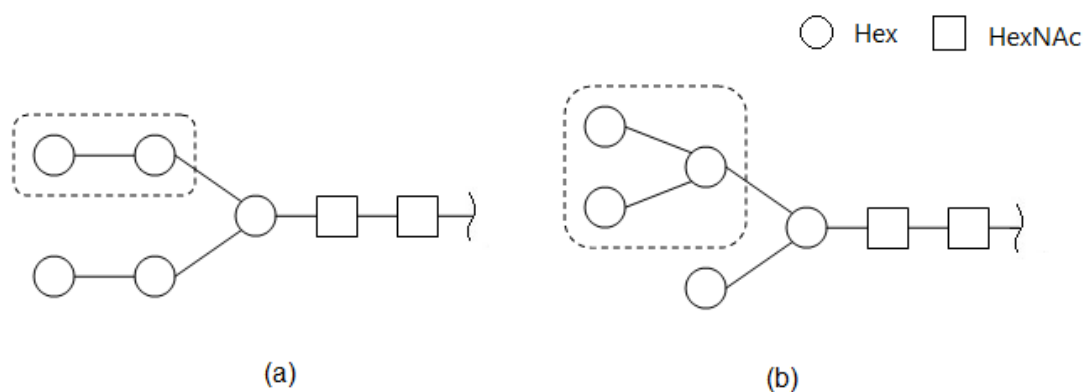


Figure 4.4: Another example of two glycan structures identified from the same spectrum by GlycoNovoDB and *de novo* sequencing algorithm respectively. Glycan (a) identified by GlycoNovoDB has higher score than glycan (b) which is reported by *de novo* sequencing method.

For the other four spectra of glycopeptide in the dataset, the glycan structures identified by *de novo* sequencing method have higher scores than those identified by GlycoNovoDB. Figure 4.5 shows the structures reported by the two methods for each spectrum. In the figure, each pair of glycan structures in the same row are reported for the same spectrum. The first glycan in each row is identified by *de novo* sequencing algorithm with highest score. The second glycan in each row is reported by GlycoNovoDB. By comparing the glycan structures in each row, we can find that the two glycans in the same row are only partially different. Empirically, it is hard to tell which one is the correct glycan associated to the given spectrum, but the structures reported by the *de novo* sequencing algorithm indeed have highest score according to the new scoring function. It also means that the structures constructed by the *de novo* sequencing algorithm match better with the given spectra than those selected from the database. Experts may be able to utilize some known biochemistry rules to select the better glycan structures according to the results. We believe that the results provide a good reference for the biochemists to find glycans that are not included in the current glycan database.

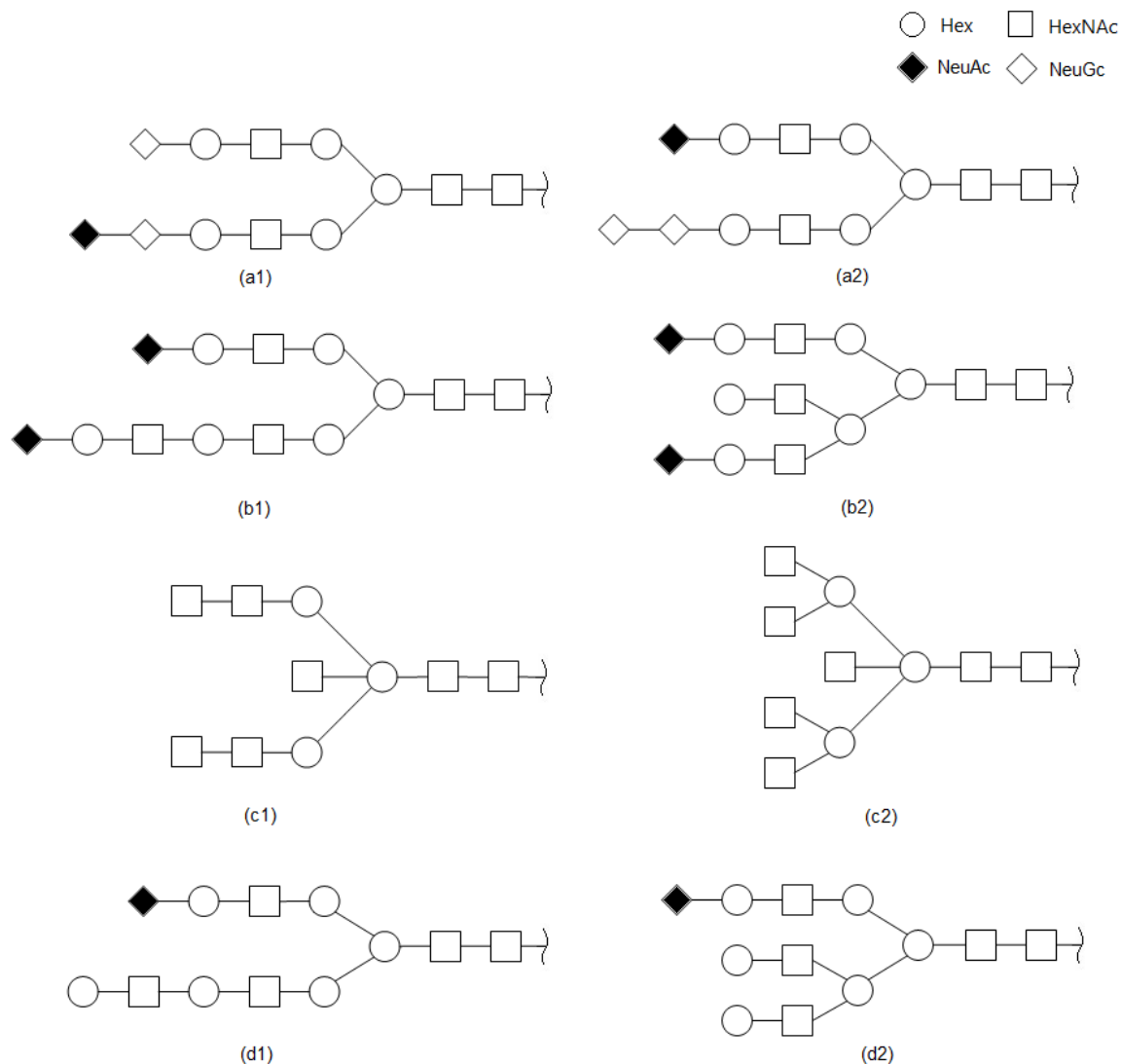


Figure 4.5: Comparison of four pairs of glycan structures in the case that glycans identified by *de novo* sequencing algorithm are ranked top according to the Function 4.10. The first glycan in each row (*i.e.* a1, b1, c1, d1) is reported by *de novo* sequencing method, while the second glycan in each row (*i.e.* a2, b2, c2, d2) is ranked top in the results reported by GlycoNovoDB.

4.5 Conclusion

In this chapter, we established a mathematical model for glycan database search problem and proposed a method for glycan structure determination from HCD glycopeptide spectra by utilizing our *de novo* sequencing results in the database search framework. The method proceeds as follows. First, the mass of peptide attached to the glycan was determined from each spec-

trum. Then we extracted all of the N-linked glycans from the glycan structure database GlycomeDB and generate glycan candidates that satisfy the mass constraint. Moreover, the glycan candidates were filtered by incorporating the *de novo* sequencing results obtained from our previously proposed algorithm. The algorithm for calculating editing based distance between unordered labeled trees were used to compute the similarity between two glycan structures from database and *de novo* sequencing results respectively.

The proposed approach was implemented as a software program GlycoNovoDB. The performance of GlycoNovoDB was compared with the software GlycoMaster DB and our previously designed *de novo* sequencing algorithm by identifying glycan structures from 46 HCD tandem mass spectra of N-linked glycopeptides. Experimental results showed that our new proposed method has higher accuracy rate than *de novo* sequencing method. Besides, GlycoNovoDB provides more meaningful results than GlycoMaster DB, since it can distinguish the multiple glycans reported by GlycoMaster DB with the same highest score more efficiently.

To take the advantages of *de novo* sequencing method and database search method efficiently, we further combined the two methods together. The improved method is to identify glycans that are in the database with high accuracy and provide new glycans that are not in the database in some cases. In the improved method, the two sets of glycan candidates obtained from *de novo* sequencing algorithm and GlycoNovoDB were merged together and compared according to their scores calculated by a single scoring function. Experimental results showed that the improved method has the ability to determine glycan structures from mass spectra with high confidence. In the case that the true glycan structure may be not included in the current database, this method can also provide a meaningful structure as a good reference.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Glycan characterization is an important topic in glycoproteomics research. It provides essential information to study the function of glycoproteins. With the development of mass spectrometers, mass spectrometry has gradually become one of the dominant techniques for determining glycan and glycopeptide primary structures. Interpreting massive and various mass spectral data manually is time consuming and tedious. Therefore it is necessary to develop software packages to automate the identification of glycan structures from mass spectral data. In the past, due to the mass range limitation, peptides and their attached glycans are usually separated before mass spectrometry analysis. In this kind of approach, glycans and peptides are analyzed individually, and the information from glycosylation sites is discarded. Alternatively, a different approach is to characterize intact glycopeptides by tandem mass spectrometry. Analysis of intact glycopeptides is a promising method for glycan and peptide identification as well as glycosylation site determination, since the glycosylation site information is conserved in the intact glycopeptide. Compared with traditional peptide sequencing from MS/MS, glycan identification is a much more challenging problem. The primary reason is that the structure of glycans is two-dimensional and the biosynthesis of glycans is a non-template driven process. There is

currently an increasing necessity for developing effective computational methods for the characterization of glycan from mass spectra. In this thesis, we conducted research regarding the identification of glycan structures from tandem mass spectra on two different but correlated topics: glycan *de novo* sequencing, and glycan identification combining *de novo* sequencing with database search.

Nowadays, the accuracy and resolution of the instruments is getting better, and there is great flexibility to choose different techniques in an experiment. Thus, it is theoretically possible to design more efficient algorithm to improve the accuracy of *de novo* sequencing. This inspired us to design a novel heuristic algorithm for glycan identification from HCD spectra, which described in Chapter 3. A mathematical model for the glycan *de novo* sequencing problem is represented as a tree construction problem. According to the algorithm, each glycan tree is constructed from root (peptide) to leaves in a heuristic manner by utilizing the quality peaks in the higher mass end of the mass spectra. A re-evaluation scheme is also applied to filter the candidate glycans to reduce the examination space of candidates. The comparison between our *de novo* sequencing algorithm and the database search method, GlycoMaster DB, shows that our proposed algorithm can effectively identify glycans from MS/MS, and its performance is comparable with GlycoMaster DB with regarding to accuracy.

Compared with the *de novo* sequencing method, database search method has the possibility to identify glycans from MS/MS with higher accuracy. Conventionally, database search is quite a successful approach for peptide sequencing. However, it is not so effective when applying to glycan identification, since the glycan databases are not quite complete yet. The algorithm proposed in Chapter 3 provides some partially correct glycan structures, and these partially correct results can be applied in database search method to assist the filtration of tentative candidates. The methods proposed to determine glycan structures in Chapter 4 are executed based on combining database search and *de novo* sequencing together. We establish a mathematical model for the glycan database search problem and propose GlycoNovoDB for

glycan identification by integrating the preliminary *de novo* sequencing results in the database search framework. The *de novo* sequencing results provide a way to score and rank the glycan candidates extracted from the glycan database. The algorithm for calculating editing based distance between unordered labelled trees is used to compute the similarity between two glycan structures from the database and *de novo* sequencing results respectively. Experiments shows that GlycoNovoDB can achieve a higher accuracy rate than the *de novo* sequencing method shown in Chapter 3, and provide more meaningful results than GlycoMaster DB. In order to identify glycans that are in the database with high accuracy as well as provide new glycans that are not in the database with confidence, we further combine database search method and *de novo* sequencing method together. In the improved method, the glycan candidates obtained from the *de novo* sequencing algorithm and GlycoNovoDB are merged together and ranked according to a revised scoring function. Experimental results show that the improved method can provide some meaningful structures as reference when the true glycan structures are not included in the database.

5.2 Future Work

The research on the characterization of glycan structures with tandem mass spectrometry remains challenging even with the progress we achieved in this research. The proteomics community will have to continue to work on the problem before we are able to discover the underlying knowledge. For the general area of glycoproteomics, our future work will be focusing on providing effective algorithmic solutions for the computational challenges described in the following section.

First, in this research we have proposed two approaches for the purpose of glycan structure determination from HCD mass spectra which includes glycan structure *de novo* sequencing and *de novo* assisted database search. However, for the *de novo* sequencing method, we

have observed that some of the top ranked structures do not conform to the biochemical rules. Under such circumstances, we can adopt a non-computational method which is to consult the researchers with biochemistry expertise to help determining the final decision of choosing the structure with the most likelihood, while another possible computational method that may be useful is to fully utilize the structural information from the glycan structure database. As we have established in our model, the glycan structure is treated as a tree and each node in the tree is constrained to have a maximum number of four children. In our previous method, we actually neglected the importance of the connection tendency between a pair of monosaccharides, and we believe the fact that one type of monosaccharide tends to connect to another type of monosaccharide with different possibilities due to the underlying biochemical properties. A feasible way to deal with the connection tendency is to make a statistical calculation based on the given glycan structure database in order to find the connecting frequencies (or probabilities) between each pair of monosaccharide residues. Such statistical information can form a matrix similar to that of the BLOSSUM matrix of amino acids which can be used to assist the process of reconstructing the glycan structure or to re-rank the structures reported with the same score by the algorithm.

Second, in the conventional shotgun proteomics research for peptide identification, a primary task the researcher is faced with is to distinguish incorrect from correct identification results after they are reported. Similarly, in our current problem of interpreting MS/MS spectra of glycopeptides, effective validation methods for estimating the incorrect glycan structures are indeed necessary for the successful analysis of large-volume glycoproteomics data. Nowadays, the Target-Decoy Database method is widely used to validate the results by providing statistical estimations on the False Discovery Rate (FDR) at a certain score threshold. For the purpose of identifying the glycan structures from database, we can apply a similar target-decoy strategy for estimating the level of uncertainty in our reported results. Methods for such a purpose are considered to be necessary especially in the case that large-scale mass spectral datasets are used

as an input and manual examination of individual glycan-structure matches (GSM) become impractical. The decoy glycan structure database being constructed should have a similar number of glycan structures, similar glycan structure mass distribution, and similar monosaccharide residue distribution as the target glycan structure database. A possible and straightforward way of generating such a decoy structure database is to label each of the glycan structures in the target database in some kind of order, for instance post order, and then reverse or shuffle the order of all the monosaccharides to form a new glycan with the same kind of tree structure. The newly generated structure will have the same number of monosaccharides, the same theoretical mass values, and more importantly, will generate the same number of subtree structures as the original target structure. After constructing the decoy glycan database, the next step we will do is to search the input MS/MS spectrum against both the target database and the decoy database to determine the likelihood of the correctness of the reported result. A general guideline is that if for a specific input spectrum, the top ranked glycan structure from the target database has a smaller score than the top ranked glycan structure from the decoy database, then we will count it as a false discovery. In the final step, only the structures from target dataset with scores above a specifically determined threshold will be reported as the trustworthy results.

Bibliography

- [1] A. Rolf, H. Henning, and S. Nathan, “On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database,” *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1473, no. 1, pp. 4–8, 1999.
- [2] R. A. Dwek, “Glycobiology: toward understanding the function of sugars,” *Chemical Reviews*, vol. 96, no. 2, pp. 683–720, 1996.
- [3] R. A. Dwek, T. D. Butters, F. M. Platt, and N. Zitzmann, “Targeting glycosylation as a therapeutic approach,” *Nature Reviews Drug Discovery*, vol. 1, no. 1, pp. 65–75, 2002.
- [4] A. Wright and S. L. Morrison, “Effect of glycosylation on antibody function: implications for genetic engineering,” *Trends in biotechnology*, vol. 15, no. 1, pp. 26–32, 1997.
- [5] K. Ohtsubo and J. D. Marth, “Glycosylation in cellular mechanisms of health and disease,” *Cell*, vol. 126, no. 5, pp. 855–867, 2006.
- [6] B. Ma, “Challenges in computational analysis of mass spectrometry data for proteomics,” *Journal of Computer Science and Technology*, vol. 25, no. 1, pp. 107–123, 2010.
- [7] A. Mayampurath, C.-Y. Yu, E. Song, J. Balan, Y. Mechref, and H. Tang, “Computational framework for identification of intact glycopeptides in complex samples,” *Analytical Chemistry*, vol. 86, no. 1, pp. 453–463, 2013.
- [8] S. Pan, R. Chen, R. Aebersold, and T. A. Brentnall, “Mass spectrometry based

- glycoproteomics—from a proteomics perspective,” *Molecular & Cellular Proteomics*, vol. 10, no. 1, p. R110.003251, 2011.
- [9] F. Li, O. V. Glinskii, and V. V. Glinsky, “Glycobioinformatics: Current strategies and tools for data mining in MS-based glycoproteomics,” *Proteomics*, vol. 13, no. 2, pp. 341–354, 2013.
- [10] I. M. Lazar, A. C. Lazar, D. F. Cortes, and J. L. Kabulski, “Recent advances in the MS analysis of glycoproteins: Theoretical considerations,” *Electrophoresis*, vol. 32, no. 1, pp. 3–13, 2011.
- [11] K. K. Lohmann and C.-W. von der Lieth, “GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates,” *Nucleic Acids Research*, vol. 32, no. suppl 2, pp. W261–W266, 2004.
- [12] E. P. Go, K. R. Rebecchi, D. S. Dalpathado, M. L. Bandu, Y. Zhang, and H. Desaire, “GlycoPep DB: a tool for glycopeptide analysis using a smart search,” *Analytical Chemistry*, vol. 79, no. 4, pp. 1708–1713, 2007.
- [13] J. M. Ren, T. Rejtar, L. Li, and B. L. Karger, “N-Glycan structure annotation of glycopeptides using a linearized glycan structure database (GlyDB),” *Journal of Proteome Research*, vol. 6, no. 8, pp. 3162–3173, 2007.
- [14] J. Albanese, M. Glueckmann, and C. Lenz, “SimGlycan™ Software*: a new predictive carbohydrate analysis tool for MS/MS data,” *Applied Biosystems*, 2010.
- [15] P. Pompach, K. B. Chandler, R. Lan, N. Edwards, and R. Goldman, “Semi-automated identification of N-Glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC–MS/MS, and glycan database search,” *Journal of proteome research*, vol. 11, no. 3, pp. 1728–1740, 2012.

- [16] L. He, L. Xin, B. Shan, G. A. Lajoie, and B. Ma, “GlycoMaster DB: software to assist the automated identification of N-linked glycopeptides by tandem mass spectrometry,” *Journal of Proteome Research*, vol. 13, no. 9, pp. 3881–389, 2014.
- [17] K.-S. Lynn, C.-C. Chen, T. M. Lih, Cheng-WeiCheng, W.-C. Su, C.-H. Chang, C.-Y. Cheng, W.-L. Hsu, Y.-J. Chen, and T.-Y. Sung, “MAGIC: an automated N-linked glyco-protein identification tool using a Y1-ion pattern matching algorithm and in silico MS2 approach,” *Analytical Chemistry*, vol. 87, no. 4, pp. 2466–2473, 2015.
- [18] H. Tang, Y. Mechref, and M. V. Novotny, “Automated interpretation of MS/MS spectra of oligosaccharides,” *Bioinformatics*, vol. 21, no. suppl.1, pp. i431–i439, 2005.
- [19] D. Goldberg, M. Bern, S. Parry, M. Sutton-Smith, M. Panico, H. R. Morris, and A. Dell, “Automated N-glycopeptide identification using a combination of single-and tandem-ms,” *Journal of Proteome Research*, vol. 6, no. 10, pp. 3995–4005, 2007.
- [20] B. Shan, B. Ma, K. Zhang, and G. Lajoie, “Complexities and algorithms for glycan sequencing using tandem mass spectrometry,” *Journal of Bioinformatics and Computational Biology*, vol. 6, no. 01, pp. 77–91, 2008.
- [21] L. Dong, B. Shi, G. Tian, Y. Li, B. Wang, and M. Zhou, “An accurate de novo algorithm for glycan topology determination from mass spectra,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 12, no. 3, pp. 568–578, 2015.
- [22] S. Böcker, B. Kehr, and F. Rasche, “Determination of glycan structure from tandem mass spectra,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 8, no. 4, pp. 976–986, 2011.
- [23] W. Sun, G. A. Lajoie, B. Ma, and K. Zhang, “A novel algorithm for glycan *de novo* sequencing using tandem mass spectrometry,” in *Bioinformatics Research and Applications*, (Norfolk, USA), pp. 320–330, June 2015.

- [24] W. Sun, M. Kuljanin, P. Pittock, B. Ma, K. Zhang, and G. A. Lajoie, “An effective approach for glycan structure *de novo* sequencing from HCD spectra,” *IEEE transactions on nanobioscience*, vol. 15, no. 2, pp. 177–184, 2016.
- [25] R. H. Garrett and C. M. Grisham, *Biochemistry*. Philadelphia: Saunders College, 1995.
- [26] Y. Gavel and G. von Heijne, “Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering,” *Protein Engineering*, vol. 3, no. 5, pp. 433–442, 1990.
- [27] N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft, and S. Brunak, “Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence,” *Proteomics*, vol. 4, no. 6, pp. 1633–1649, 2004.
- [28] M. T. Davis and T. D. Lee, “Rapid protein identification using a microscale electrospray LC/MS system on an ion trap mass spectrometer,” *Journal of the American Society for Mass Spectrometry*, vol. 9, no. 3, pp. 194–201, 1998.
- [29] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, “Probability-based protein identification by searching sequence databases using mass spectrometry data,” *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.
- [30] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, “PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry,” *Rapid communications in mass spectrometry*, vol. 17, no. 20, pp. 2337–2342, 2003.
- [31] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold, “Quantitative analysis of complex protein mixtures using isotope-coded affinity tags,” *Nature biotechnology*, vol. 17, no. 10, pp. 994–999, 1999.
- [32] S.-E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann, “Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and

- accurate approach to expression proteomics,” *Molecular & cellular proteomics*, vol. 1, no. 5, pp. 376–386, 2002.
- [33] W. Wang, H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, M. Anderle, and C. H. Becker, “Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards,” *Analytical chemistry*, vol. 75, no. 18, pp. 4818–4826, 2003.
- [34] H. Lin, L. He, and B. Ma, “A combinatorial approach to the peptide feature matching problem for label-free quantification,” *Bioinformatics*, vol. 29, no. 14, pp. 1768–1775, 2013.
- [35] R. Aebersold and M. Mann, “Mass spectrometry-based proteomics,” *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.
- [36] M. Karas and F. Hillenkamp, “Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons,” *Analytical chemistry*, vol. 60, no. 20, pp. 2299–2301, 1988.
- [37] F. Hillenkamp, M. Karas, R. C. Beavis, and B. T. Chait, “Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers,” *Analytical chemistry*, vol. 63, no. 24, pp. 1193A–1203A, 1991.
- [38] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, “Electrospray ionization for mass spectrometry of large biomolecules,” *Science*, vol. 246, no. 4926, pp. 64–71, 1989.
- [39] M. Mann and M. Wilm, “Electrospray mass spectrometry for protein characterization,” *Trends in biochemical sciences*, vol. 20, no. 6, pp. 219–224, 1995.
- [40] A. D. McNaught and A. D. McNaught, *Compendium of chemical terminology*, vol. 1669. Blackwell Science Oxford, 1997.

- [41] P. Dawson, "Quadrupole mass analyzers: performance, design and some recent applications," *Mass Spectrometry Reviews*, vol. 5, no. 1, pp. 1–37, 1986.
- [42] K. R. Jonscher and J. R. Yates, "The quadrupole ion trap mass spectrometer a small solution to a big challenge," *Analytical biochemistry*, vol. 244, no. 1, pp. 1–15, 1997.
- [43] J. W. Hager, "A new linear ion trap mass spectrometer," *Rapid Communications in Mass Spectrometry*, vol. 16, no. 6, pp. 512–526, 2002.
- [44] B. Mamyrin, "Time-of-flight mass spectrometry (concepts, achievements, and prospects)," *International Journal of Mass Spectrometry*, vol. 206, no. 3, pp. 251–266, 2001.
- [45] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson, "Fourier transform ion cyclotron resonance mass spectrometry: a primer," *Mass spectrometry reviews*, vol. 17, no. 1, pp. 1–35, 1998.
- [46] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks, "The Orbitrap: a new mass spectrometer," *Journal of mass spectrometry*, vol. 40, no. 4, pp. 430–443, 2005.
- [47] C. Barner-Kowollik, T. Gruendling, J. Falkenhagen, and S. Weidner, *Mass spectrometry in polymer chemistry*. John Wiley & Sons, 2012.
- [48] L. He, "Algorithms for characterizing peptides and glycopeptides with mass spectrometry," 2013.
- [49] K. Tang, J. S. Page, and R. D. Smith, "Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry," *Journal of the American Society for Mass Spectrometry*, vol. 15, no. 10, pp. 1416–1423, 2004.
- [50] J. M. Wells and S. A. McLuckey, "Collision-induced dissociation (CID) of peptides and proteins," *Methods in enzymology*, vol. 402, pp. 148–185, 2005.

- [51] J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9528–9533, 2004.
- [52] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann, "Higher-energy C-trap dissociation for peptide modification analysis," *Nature methods*, vol. 4, no. 9, pp. 709–712, 2007.
- [53] M. Wührer, M. I. Catalina, A. M. Deelder, and C. H. Hokke, "Glycoproteomics based on tandem mass spectrometry of glycopeptides," *Journal of Chromatography B*, vol. 849, no. 1, pp. 115–128, 2007.
- [54] A. L. Burlingame, "Characterization of protein glycosylation by mass spectrometry," *Current opinion in biotechnology*, vol. 7, no. 1, pp. 4–10, 1996.
- [55] D. J. Harvey, "Proteomic analysis of glycosylation: structural determination of N-and O-linked glycans by mass spectrometry," *Expert Review of Proteomics*, vol. 2, no. 1, pp. 87–101, 2005.
- [56] D. S. Dalpathado and H. Desaire, "Glycopeptide analysis by mass spectrometry," *Analyst*, vol. 133, no. 6, pp. 731–738, 2008.
- [57] C. L. Woodin, M. Maxon, and H. Desaire, "Software for automated interpretation of mass spectrometry data from glycans and glycopeptides," *Analyst*, vol. 138, no. 10, pp. 2793–2803, 2013.
- [58] A. Marimuthu, R. N. O'Meally, R. Chaerkady, Y. Subbannayya, V. Nanjappa, P. Kumar, D. S. Kelkar, S. M. Pinto, R. Sharma, S. Renuse, *et al.*, "A comprehensive map of the human urinary proteome," *Journal of proteome research*, vol. 10, no. 6, pp. 2734–2743, 2011.

- [59] P. Roepstorff and J. Fohlman, "Letter to the editors," *Biological Mass Spectrometry*, vol. 11, no. 11, pp. 601–601, 1984.
- [60] R. S. Johnson, S. A. Martin, K. Biemann, J. T. Stults, and J. T. Watson, "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine," *Analytical chemistry*, vol. 59, no. 21, pp. 2621–2625, 1987.
- [61] J. Zaia, "Mass spectrometry of oligosaccharides," *Mass Spectrometry Reviews*, vol. 23, no. 3, pp. 161–227, 2004.
- [62] C. K. Frese, A. M. Altelaar, M. L. Hennrich, D. Nolting, M. Zeller, J. Griep-Raming, A. J. Heck, and S. Mohammed, "Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos," *Journal of Proteome Research*, vol. 10, no. 5, pp. 2377–2388, 2011.
- [63] W. R. Alley, Y. Mechref, and M. V. Novotny, "Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data," *Rapid Communications in Mass Spectrometry*, vol. 23, no. 1, pp. 161–170, 2009.
- [64] N. E. Scott, B. L. Parker, A. M. Connolly, J. Paulech, A. V. Edwards, B. Crossett, L. Falconer, D. Kolarich, S. P. Djordjevic, P. Højrup, *et al.*, "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*, vol. 10, no. 2, pp. M000031–MCP201, 2011.
- [65] C. Singh, C. G. Zampronio, A. J. Creese, and H. 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*, vol. 11, no. 9, pp. 4517–4525, 2012.
- [66] S. Doubet, K. Bock, D. Smith, A. Darvill, P. Albersheim, and P. Albersheim are at the University of Georgia, “The complex carbohydrate structure database,” *Trends in biochemical sciences*, vol. 14, no. 12, pp. 475–477, 1989.
- [67] A. Loß, P. Bunsmann, A. Bohne, A. Loß, E. Schwarzer, E. Lang, and C.-W. von der Lieth, “SWEET-DB: an attempt to create annotated data collections for carbohydrates,” *Nucleic acids research*, vol. 30, no. 1, pp. 405–408, 2002.
- [68] T. Lütteke, A. Bohne-Lang, A. Loss, T. Goetz, M. Frank, and C.-W. von der Lieth, “GLY-COSCIENCES. de: an internet portal to support glycomics and glycobiology research,” *Glycobiology*, vol. 16, no. 5, pp. 71R–81R, 2006.
- [69] C.-W. von der Lieth, A. A. Freire, D. Blank, M. P. Campbell, A. Ceroni, D. R. Damerell, A. Dell, R. A. Dwek, B. Ernst, R. Fogh, *et al.*, “EUROCarbDB: an open-access platform for glycoinformatics,” *Glycobiology*, vol. 21, no. 4, pp. 493–502, 2011.
- [70] C. A. Cooper, M. J. Harrison, M. R. Wilkins, and N. H. Packer, “GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources,” *Nucleic Acids Research*, vol. 29, no. 1, pp. 332–335, 2001.
- [71] C. A. Cooper, H. J. Joshi, M. J. Harrison, M. R. Wilkins, and N. H. Packer, “GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update,” *Nucleic acids research*, vol. 31, no. 1, pp. 511–513, 2003.
- [72] R. Ranzinger, S. Herget, T. Wetter, and C.-W. Von Der Lieth, “GlycomeDB—integration of open-access carbohydrate structure databases,” *BMC bioinformatics*, vol. 9, no. 1, p. 1, 2008.

- [73] R. Ranzinger, S. Herget, C.-W. von der Lieth, and M. Frank, “GlycomeDB—a unified database for carbohydrate structures,” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D373–D376, 2011.
- [74] K. K. Lohmann and C.-W. von der Lieth, “Glyco-fragment: A web tool to support the interpretation of mass spectra of complex carbohydrates,” *Proteomics*, vol. 3, no. 10, pp. 2028–2035, 2003.
- [75] H. J. Joshi, M. J. Harrison, B. L. Schulz, C. A. Cooper, N. H. Packer, and N. G. Karlsson, “Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data,” *Proteomics*, vol. 4, no. 6, pp. 1650–1664, 2004.
- [76] A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell, and S. M. Haslam, “GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans,” *Journal of proteome research*, vol. 7, no. 4, pp. 1650–1659, 2008.
- [77] D. J. Pappin, P. Hojrup, and A. J. Bleasby, “Rapid identification of proteins by peptide-mass fingerprinting,” *Current biology*, vol. 3, no. 6, pp. 327–332, 1993.
- [78] J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. A. Lajoie, and B. Ma, “PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification,” *Molecular & Cellular Proteomics*, vol. 11, no. 4, pp. M111–010587, 2012.
- [79] J. K. Eng, A. L. McCormack, and J. R. Yates, “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database,” *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.
- [80] R. Craig and R. C. Beavis, “TANDEM: matching proteins with tandem mass spectra,” *Bioinformatics*, vol. 20, no. 9, pp. 1466–1467, 2004.

- [81] S. P. Gaucher, J. Morrow, and J. A. Leary, “STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry,” *Analytical chemistry*, vol. 72, no. 11, pp. 2331–2336, 2000.
- [82] M. Ethier, J. A. Saba, W. Ens, K. G. Standing, and H. Perreault, “Automated structural assignment of derivatized complex N-linked oligosaccharides from tandem mass spectra,” *Rapid communications in mass spectrometry*, vol. 16, no. 18, pp. 1743–1754, 2002.
- [83] S. Kumozaki, K. Sato, and Y. Sakakibara, “A machine learning based approach to *de novo* sequencing of glycans from tandem mass spectrometry spectrum,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 6, pp. 1267–1274, 2015.
- [84] A. P. Snyder *et al.*, *Interpreting protein mass spectra*. American Chemical Society; Oxford University Press, 2000.
- [85] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*. Reading, Massachusetts, USA: Addison-Wesley Publishing Company, 1974.
- [86] M. J. Huddleston, M. F. Bean, and S. A. Carr, “Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: methods for selective detection of glycopeptides in protein digests,” *Analytical chemistry*, vol. 65, no. 7, pp. 877–884, 1993.
- [87] K. Zhang, R. Statman, and D. Shasha, “On the editing distance between unordered labeled trees,” *Information processing letters*, vol. 42, no. 3, pp. 133–139, 1992.
- [88] K. Zhang, “A new editing based distance between unordered labeled trees,” in *Annual Symposium on Combinatorial Pattern Matching*, pp. 254–265, Springer, 1993.
- [89] K. Zhang, “A constrained edit distance between unordered labeled trees,” *Algorithmica*, vol. 15, no. 3, pp. 205–222, 1996.
- [90] S. Chen, B. Ma, and K. Zhang, “On the similarity metric and the distance metric,” *Theoretical Computer Science*, vol. 410, no. 24, pp. 2365–2376, 2009.

Curriculum Vitae

Name: Weiping Sun

Post-Secondary Education and Degrees: Central South University
Changsha, Hunan Province, P.R. China
2009-2012 M. Eng.

Central South University
Changsha, Hunan Province, P.R. China
2005-2009 Honours. B. Eng.

Honours and Awards: Western Graduate Research Scholarship
2012-2016

Academic Training Teaching & Research Assistant
The University of Western Ontario
2012-2016

Publications:

Thesis-related publications:

1. Weiping Sun, Gilles A. Lajoie, Bin Ma, and Kaizhong Zhang. “A Novel Algorithm for Glycan *De Novo* Sequencing Using Tandem Mass Spectrometry.” In *Proceedings of 11th International Symposium on Bioinformatics Research and Applications (ISBRA 2015)*, pp. 320-330, 2015.

Authors’ contribution: WS and KZ conceived the study. WS wrote the manuscript and code. GAL prepared the experimental data. BM participated in the discussion and provided comments. KZ revised the manuscript and provide suggestions.

2. Weiping Sun, Miljan Kuljanin, Paula Pittock, Bin Ma, Kaizhong Zhang, and Gilles A. Lajoie. “An Effective Approach for Glycan Structure *De Novo* Sequencing From HCD Spectra.” *IEEE Transactions on Nanobioscience*, 15(2):177-184, March 2016.

Authors’ contribution: WS and KZ conceived the study. WS wrote the manuscript and code. MK, PP, and GAL prepared the experimental data. BM participated in the discussion and provided comments. KZ revised the manuscript and provide suggestions.

3. Weiping Sun, Yi Liu, Gilles A. Lajoie, Bin Ma, and Kaizhong Zhang. “An Improved Approach for N-linked Glycan Structure Identification from HCD MS/MS Spectra.” (Accepted by *the 27th International Conference on Genome Informatics (GIW 2016)*, and invited to submit to *IEEE Transactions on Computational Biology and Bioinformatics*).

Authors’ contribution: WS and KZ conceived the study. WS wrote the paper and code. GAL prepared the experimental data. YL and BM participated in the discussion and provided comments. KZ revised the manuscript and provide suggestions.

Thesis-unrelated publications:

1. Yi Liu, Weiping Sun, Gilles A. Lajoie, Bin Ma, and Kaizhong Zhang. “An Approach for Matching Mixture MS/MS Spectra with a Pair of Peptide Sequences in a Protein Database.” In *Proceedings of 11th International Symposium on Bioinformatics Research and Applications (ISBRA 2015)*, pp. 223-234, 2015.

2. Yi Liu, Weiping Sun, Julius John, Gilles A. Lajoie, Bin Ma, and Kaizhong Zhang. “De Novo Sequencing Assisted Approach for Characterizing Mixture MS/MS Spectra.” IEEE Transactions on Nanobioscience 15(2): 166-176, March 2016.