2015

De novo sequencing of heparan sulfate saccharides using high-resolution tandem mass spectrometry

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

DE NOVO SEQUENCING OF HEPARAN SULFATE SACCHARIDES USING HIGH-RESOLUTION TANDEM MASS SPECTROMETRY

by

HAN HU

B.A., Sichuan University, China, 2007

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2015

Approved by

First Reader _____

Joseph Zaia, Ph.D.
Professor of Biochemistry

Second Reader _____

Yu (Brandon) Xia, Ph.D.
Adjunct Associate Professor of Bioinformatics, Boston University
Associate Professor of Bioengineering, McGill University, Canada

DEDICATION

I would like to dedicate this work to my parents, my aunt, and all my friends who accompanied me on my PhD journey.

ACKNOWLEDGMENTS

Since the day when I received a late admission offer from BU, my PhD journey has never become easier. I was so surprised to see the twists and turns queuing in my way and it was so difficult to get through on my own. It's a great fortune to meet so many people whom I can trust and rely on. Without their support and advice, I cannot insist till the graduation of my PhD.

My deepest gratitude goes to my advisor, Dr. Joseph Zaia, for his continuous support throughout my research. When I started my rotation with Joe, I just left my old lab and was anxiously looking for places to continue my research career. Joe accepted me and gave me the freedom with trust so that I can have enough time to explore the research projects, to try, fail and improve any new methods that came to my mind. No matter technical problem or career guidance, Joe was always patient and ready to support.

My deepest gratitude also goes to my co-advisor, Dr. Yu Xia, for his inspiring ideas and encouraging visions of science, research and life. I met Brandon on my first semester at BU, and worked under his guidance for my second rotation. There were many valuable suggestions that I could not catch on time until several years later.

Special thanks to my committee members: Dr. Scott Mohr, Dr. Mark Kon and Dr. Mark McComb for their advice during my research. They gave me a lot of encouragement and support for me to finish my thesis project.

My sincere appreciation goes to Dr. Scott Mohr, Dr. Tom Tullius, Dave, Caroline, Johanna, and all people from our Bioinformatics Program. When I had trouble with switching lab, when I had problem with my visa and student status, I felt so desperate and

lost my direction. People in our program supported me in those dark periods and encouraged me to continue my research. There are so many times I felt like I was in a big family, and I will keep that in mind for the rest of my life.

I also want to extend my deepest gratitude to my family. When I was preparing for my thesis defense, my mother had a car accident. My parents and all my relatives concealed that from me in order to have me focus on my dissertation. I owe a lot to them. Without their encouragement and support, I would not stick it out.

DE NOVO SEQUENCING OF HEPARAN SULFATE SACCHARIDES USING HIGH-RESOLUTION TANDEM MASS SPECTROMETRY

(Order No.)

HAN HU

Boston University Graduate School of Arts and Sciences

and College of Engineering, 2015

Major Professor: Joseph Zaia, Professor of Biochemistry

ABSTRACT

Heparan sulfate (HS) is a class of linear, sulfated polysaccharides located on cell surface, secretory granules, and in extracellular matrices found in all animal organ systems. It consists of alternately repeating disaccharide units, expressed in animal species ranging from hydra to higher vertebrates including humans. HS binds and mediates the biological activities of over 300 proteins, including growth factors, enzymes, chemokines, cytokines, adhesion and structural proteins, lipoproteins and amyloid proteins. The binding events largely depend on the fine structure – the arrangement of sulfate groups and other variations – on HS chains.

With the activated electron dissociation (ExD) high-resolution tandem mass spectrometry technique, researchers acquire rich structural information about the HS molecule. Using this technique, covalent bonds of the HS oligosaccharide ions are dissociated in the mass spectrometer. However, this information is complex, owing to the large number of product ions, and contains a degree of ambiguity due to the overlapping of product ion masses and lability of sulfate groups; as a result, there is a

serious barrier to manual interpretation of the spectra. The interpretation of such data creates a serious bottleneck to the understanding of the biological roles of HS. In order to solve this problem, I designed HS-SEQ – the first HS sequencing algorithm using high-resolution tandem mass spectrometry. HS-SEQ allows rapid and confident sequencing of HS chains from millions of candidate structures and I validated its performance using multiple known pure standards. In many cases, HS oligosaccharides exist as mixtures of sulfation positional isomers. I therefore designed MULTI-HS-SEQ, an extended version of HS-SEQ targeting spectra coming from more than one HS sequence. I also developed several pre-processing and post-processing modules to support the automatic identification of HS structure. These methods and tools demonstrated the capacity for large-scale HS sequencing, which should contribute to clarifying the rich information encoded by HS chains as well as developing tailored HS drugs to target a wide spectrum of diseases.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AT ..... Antithrombin

CAD ..... Collisionally activated dissociation

CID ..... Collision-induced dissociation

ECD ..... Electron-capture dissociation

ECM ..... Extracellular matrix

EDD ..... Electron-detached dissociation

EST ..... Expressed sequence tags

ETD ..... Electron-transfer dissociation

ExD ..... Activated electron dissociation

FDR ..... False discovery rate

FGF ..... Fibroblast growth factor

FGFR ..... Fibroblast growth factor receptor

FTICR ..... Fourier transform ion cyclotron resonance

GA ..... Generic algorithm

GAG ..... Glycosaminoglycan

GPI ..... Glycosphosphatidylinositol

HCD ..... High-energy collision dissociation

HS ..... Heparan sulfate

HSBP ..... Heparan sulfate binding protein

HSPG ..... Heparan sulfate proteoglycan

LC ..... Liquid chromatography

MS ..... Mass spectrometry

MVC .....Model-view-controller

NETD..... Negative electron-transfer dissociation

NMR ..... nuclear magnetic resonance

PAGE .....Polyacrylamide gel electrophoresis

PSM ..... Peptide-spectrum match

PTM ..... Post-translational modification

QC..... Quality control

S/N ..... Signal-to-noise

TDA ..... Target-decoy approach

Chapter 1 Introduction

1.1 Overview of Dissertation

The dissertation is organized as follows:

Chapter 1 introduces the background of heparan sulfate: its physiological role, chemical structure, biosynthesis and its binding mechanism with proteins.

Chapter 2 reviews the state-of-the-art identification algorithms in proteomics and glycoproteomics, and discusses the possibility of migrating the algorithms into heparan sulfate identification.

Chapter 3 presents HS-SEQ, the first de novo sequencing algorithm for identifying the sulfation pattern on heparan sulfate sequence and compares the performance of HS-SEQ with two naïve methods.

Chapter 4 presents MULTI-HS-SEQ, an expanded version of HS-SEQ in identifying heparan sulfate sulfation pattern in the context of mixture.

Chapter 5 introduces several pre-processing and post-processing modules and algorithms assisting the working of HS-SEQ.

Chapter 6 summarizes the thesis, discusses the practical significance of HS-SEQ in promoting scientific study and drug design, and explores the possibility of generalizing the model of HS-SEQ to identification of peptides and glycopeptides.

Chapter 3 has been published in Molecular Cellular & Proteomics (2014). Chapter 2 is organized as a review and publishing is under planning. Chapter 4 and 5 will be combined as a pipeline paper, whereas 5.2.2 has been published in Analytical Chemistry (2013).

1.2 Physiological roles of Heparan Sulfate

Heparan sulfate (HS) is a type of linear polysaccharides consisting of alternative repeating disaccharides belonging to glucosaminoglycan (GAG) family. HS is an ancient molecule with conserved basic structure over 500 million years of evolution. From primitive species such as hydra to higher vertebrates including humans, HS was found in all animals examined (1), with the exception of Porifera (*e.g.* sponge) (2). At the cellular level, it locates on cell surface, in extracellular matrix (ECM) as well as intracellular granules, and mediates cell-cell interaction, matrix remodeling and activation of multiple signaling pathways (Figure 1). At the tissue level, it is involved in all animal organ systems (3), and responsible for tissue development, anticoagulation, angiogenesis, wound repair, pathogen recognition and many other biological functions (3). The active roles of HS in inflammation (4) and cancer (5, 6) have also been intensively reported.

One of the most intensely studied examples of HS is its effect on anticoagulation. Heparin, the highly sulfated form of HS, binds to antithrombin (AT) III, which causes a conformational change of ATIII and eventually its activation. The heparin/ATIII complex is able to inhibit the activity of proteases (*e.g.* thrombin, factor Xa) in fibrin clot formation, and therefore prevents blood coagulation. Another well-known example is the binding between HS and fibroblast growth factors (FGFs), which forms morphogen gradients in the ECM and leads to branching morphogenesis (7). Besides, HS competes with chondroitin sulfate (CS), another GAG member, to bind to receptor protein tyrosine phosphatase sigma (RPTPσ) (8), which regulates the development and repair of nervous system.

All HS biological activities pivot on its structural specialties, which dictates the binding with diverse proteins. In fact, HS is capable of binding to > 300 proteins (heparan sulfate-binding proteins, HSBPs) (9), including growth factors, chemokines, cytokines, blood coagulation factors, structural proteins, lipoproteins and amyloid proteins, in a non-covalent manner. Pathogens can recognize HS sequence for their invasion into the host. On the other hand, HS is able to bind covalently to a relatively small set of core proteins (~17) either on cell surface (*e.g.* glypican) or in ECM (*e.g.* serglycin). The HS-core protein complex, which is named as heparan sulfate proteoglycan (HSPG), restricts the spatial distribution of HS chains, and therefore provides a framework to regulate the binding between HS and HSBPs. These facts implicate that HS either resembles the role of simple molecules like H_2O, to assist the binding of molecules in a straightforward way, or follows the style of large molecules like RNA and protein, to involve in biological functions based on definitive sequences. Surprisingly, the debate on the binding mechanism between HS and HSBPs lasts for decades, regardless of the emerging new discoveries on HS-ligand interaction (10).

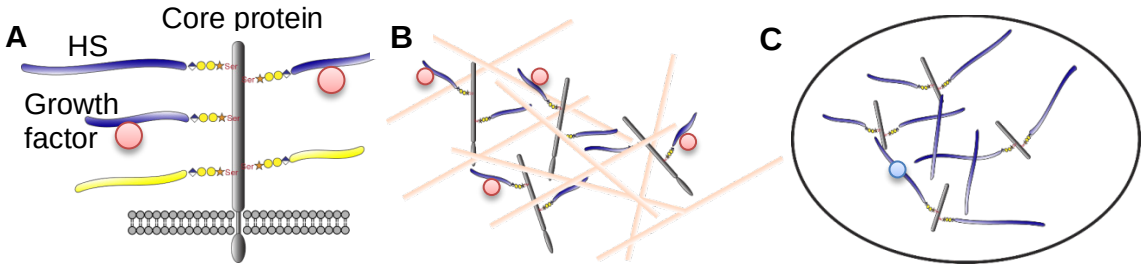


Figure 1 Heparan sulfate plays an universal role in cell phisiology. HS locates on (A) cell surface, in (B) ECM and (C) secreted granules, and interacts with multiple types of proteins.

1.3 Structure and Biological Synthesis of Heparan Sulfate Sequence

HS molecule structure is highly expressive and organized. HS chain consists of repeating disaccharide units [-4-(GlcA-β/IdoA-α)-1,4-GlcNAc(NS)-α-], where GlcA/IdoA may undergo 2-*O*-sulfation, and GlcNAc may undergo *N*-deacetylation (free -NH_2), *N*-sulfation, 6-*O*-sulfation, and in rare cases, 3-*O*-sulfation (Figure 2A). Theoretically there is 48 disaccharide variants covering different combinations of sulfation, acetylation and epimerization. However, disaccharide analysis suggested that only ~20 variants are commonly present in animal tissues, which is comparable to the 20 amino acid residues. Therefore, a HS sequence with *m* disaccharide units is equivalent to a peptide with *m* amino acid residues in terms of the sequence variations. This comparison may be unfair, since not all disaccharide units contribute to the binding activity. Instead, the disaccharide units on HS chain constitute domains characteristic of distinct sulfation degrees. A typical HS chain contains several highly sulfated domains (NS domains) interspersed with rarely sulfated domains (NA domains) and transient domains (NA/NS domains) (Figure 2B). The layout of the domains are biologically meaningful, as HSBPs frequently bind to NS domains and a short NA domain may serve to the oligomerization of protein ligands (*e.g.* chemokine). Variation also exists within NS domains. The NS domain can vary from dp4 up to dp18, accommodating the assembly of different protein ligands. Factors such as chain length, sulfation degree, sulfation/epimerization pattern and heterogeneity among sequences on the same core protein, contribute to the biological activity of HS and make HS molecule one of the most complicate biopolymer in nature.

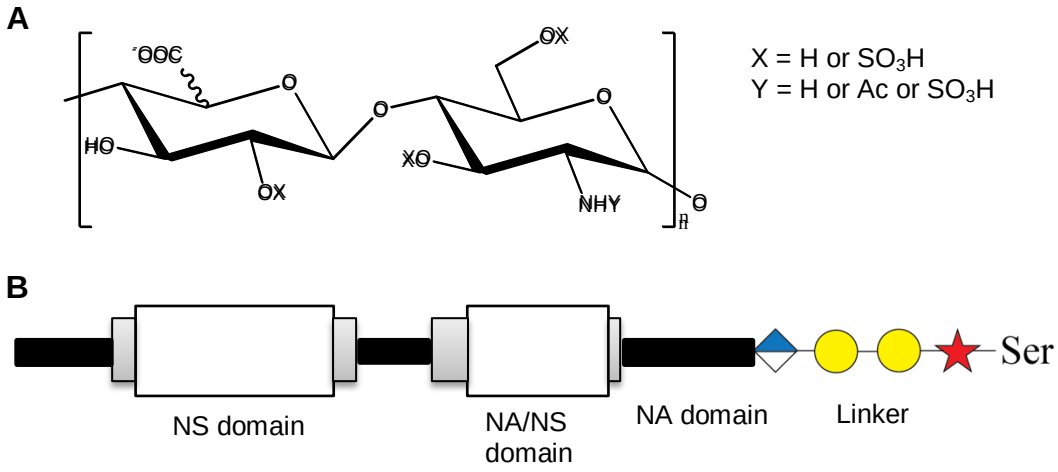


Figure 2 Heparan sulfate structure. (A) The repeating disaccharide unit of HS. (B) The domain structure of HS chain determined by sulfation degree.

The complicate structure of HS results from strictly controlled biosynthesis procedure in the Golgi apparatus and endoplasmic reticulum (ER) (11). Different from DNA, RNA and protein sequence, whose synthesis relies on precise templates, all glycan synthesis is non-template driven. How multiple enzymes coordinate with each other in a temporal and spatial manner has not been fully illustrated. A classical HS biosynthesis model consists of three steps, including chain initialization, polymerization and polymer modification. In the initialization stage, the core protein is first xylosylated by xylotransferase 1 (XYLT1) and xylotransferase 2 (XYLT2) at selected serine residue. Two galactose (Gal) residues are then successively appended to the xylose (Xyl) residue, catalyzed by galactosyltransferase 1 (GalT1) and 2 (GalT2), respectively. A glucuroinc acid (GlcA) residue is further added by glucuronosyltransferase 1 (GlcAT1) to complete

the tetrasaccharide linker. Further modifications can occur on the linker region, such as phosphorylation on Xyl and sulfation on the two Gal residues. These modifications can affect following polymerization direction, and possibly determine the bifurcation of HS and CS biosynthesis (12).

In the polymerization stage, the first GlcNAc is added to the linker by exostose-like protein 3 (EXTL3). GlcA and GlcNAc are then alternately appended to the chain through EXT1/EXT2 copolymerase. The synthesized polysaccharide further undergoes extensive modifications, including 2-*N* modification of GlcNAc (N-deacetylation/N-sulfation) through *N*-deacetylase/*N*-sulfotransferase (NDST 1 and 2), C5 epimerization of GlcA to IdoA, 2-*O* sulfation of IdoA by a 2-*O*-sulfotransferase (2OST), 6-*O* sulfation by a 6-*O*-sulfotransferase (6OST) of the glucosamine residue and, in rare cases, additional 3-*O* sulfation through 3-*O*-sulfotransferase (3OST). Note that to date there're only one 2-*O*-sulfotransferase identified, but three 6-*O*-sulfotransferases (6OST1-3) and seven 3-*O*-sulfotransferases (3OST1, 2, 3a, 3b, 4, 5, and 6) known to exist. The increasing number of sulfotransferase isozymes reflects the growing complexity and substrate diversity along the HS biosynthesis path. On the other hand, this may also implicate a large population of sulfation arrangements available on HS chain.

Synthesized HS chains can undergo further modification when they arrive at the cell surface or in the ECM. Endosulfatases SULF1 and SULF2 anchored to the cell surface can remove specific 6-O sulfate groups, and thus regulate the activity of HS binding with Wnt, BMP and FGF. Besides, heparanase is able to truncate HS chain into

short, free forms of HS oligosaccharides, which play an important role in metastasis and angiogenesis.

1.4 Heparan Sulfate and Protein Interaction

As mentioned above, HS chains connect to a specific serine residues on core protein (Figure 2B), and form a special glycoprotein complex – heparan sulfate proteoglycan (HSPG). To date, there have been ~17 core proteins reported, which can be grouped according to their respective locations (13): 1) membrane HSPGs, which includes transmembrane HSPGs, such as syndecans, and glycosylphosphatidylinositol (GPI)-anchored HSPGs such as glypican; 2) secreted ECM HSPGs, such as agrin, perlecan, collagen XVIII; and 3) secretory vesicle HSPGs, such as serglycin. Some types of proteoglycans contain only one HS chain (*e.g.* CD44v3 and betaglycan), while others may have more (*e.g.* serglycin). Proteoglycans may also contain other glycans such as chondroitin sulfate (CS) or even mucin-type *O*-glycan, which significantly expands the flexibility of HSPGs in protein binding.

In contrast with the relatively small number of core proteins, there have been over 300 proteins (heparan sulfate-binding proteins, HSBPs) identified to bind with HS in non-covalent ways (Figure 3). The HSBPs cover a wide spectrum of categories (9), including chemokines and cytokines, growth factors and morphogens, blood coagulation factors, support and structural proteins, signaling receptors, cell adhesion proteins, lipid-binding proteins and amyloids. Most HSBPs are evolutionarily unrelated to each other, yet form HS-binding sites consisting of different domain structures and bind HS with a

wide dynamic range of binding affinities (dissociation constant K_d ranges from 1 nM to 10 µM) (9). This suggests that the binding between HSBPs and HS follows convergent evolution.

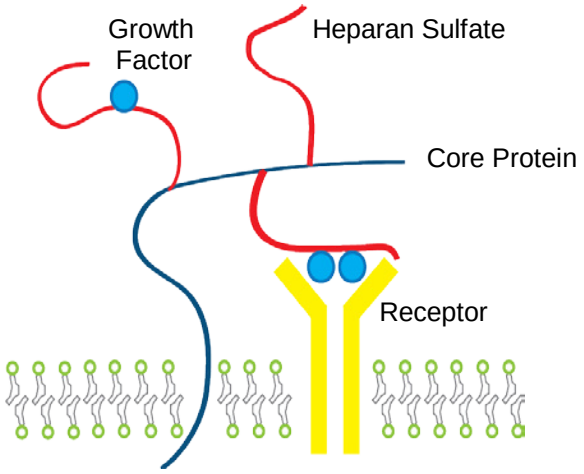


Figure 3 Heparan sulfate binds with core proteins and many other proteins. The picture shows the triplex of HS-HSBP-HSPG. HS is critical in facilitating the binding between growth factors and their receptors.

The binding between HS and HSBPs follows several distinct patterns (14). A single protein may bind to a single HS domain, such as the binding between HS and antithrombin; the binding of a protein to a single HS domain may promote the binding of its receptor to the same domain, and form a ternary complex, such as the binding between HS, FGF and FGF receptor (FGFR); two heterogeneous proteins may bind to two different HS domains and form a ternary complex, such as the binding between HS, antithrombin, and thrombin; homo-proteins such as chemokines, can bind to neighboring HS domains in *cis* or *trans* manner, and form oligomers.

The predominant way for HS to interact with HSBP is through electrostatic interaction, where the negative charged groups from HS (sulfate groups and carboxyl groups) interact with positive charged lysine and arginine residues from the protein. In this case, the density of sulfate groups dominates the binding affinity. A typical example is the binding of HS to thrombin (15), where the involved amino acid residues are capable of binding with different sets of sulfate and carboxylate groups, according to the orientations of the heparin. However, the interaction can still be selective, as many HSBPs bind with HS only through a distinct subset of lysine and arginine residues, regardless of the rest positively charged groups on the surface. Another important way of HS-HSBP binding is based on the formation of hydrogen bonds, where HS interacts with polar residues such as asparagine, glutamine and histidine. In this way, the binding between HS and HSBPs requires highly specific HS structures (*e.g.* the binding between an optimal pentasaccharide from heparin and ATIII), and receives wide interests in looking for HS motifs that dictate the binding events and mediate downstream biological functions.

1.5 Exploration of "Glycan Code" on Heparan Sulfate Chains

The strictly controlled biosynthesis procedure, information-rich structure and intriguing protein binding mechanisms of HS motivated the search of "glycan code" (16) – specific arrangements of modifications (sulfation and epimerization) on HS that binds protein ligands with high specificity. There have been years of debate regarding the existence of HS motifs which could uniquely bind proteins.

Recently, advances in other GAG family members ignited the enthusiasm for deterministic GAG sequences. Ly *et al.* (17) sequenced bikunin, the simplest preoteoglycan with a single CS chain attached, using top-down FTMS analysis, and showed that it possesses a single or small number of defined sequences regardless the overall chain length. Zhao *et al.* (18) applied similar methods and drew the same conclusion for decorin, another proteoglycan with a more complicate DS/CS GAG chain attached. Given the background that all GAG members share common biosynthetic pathway, these studies strongly suggest that HS, one of the most complicate and expressive biopolymers, may also contain deterministic motifs that bind with protein ligands and play special biological roles.

Due to the paucity of available HS samples (milligrams), most HS studies relied on the derivation or degradation of heparin (kilograms) by assuming that the dense sulfation on heparin chain could effectively represent the true HS sequence that binds to target protein ligand, and surrounding sulfate groups from heparin products would not interfere the binding (9). This assumption may not hold, given the capacity of diverse HS sequences as well as the observation that misplaced sulfate groups even reduced the binding affinity between heparin oligosaccharides and ligands.

Studies of sulfation prevalence in natural heparan sulfate revealed that 3-*O*-sulfation constitutes the smallest population. As a comparison, *hs3st*s represent the largest gene family (7 members identified in vertebrates) in HS sulfotransferases and isozymes of *hs3st*s are differentially and widely spread in animal tissues (19). 3-*O*-sulfation is known to boost the binding affinity between ATIII and heparin by ~ 10^5

orders of magnitude, and preside the recognition of glycoprotein gD of type I herpes simplex virus. Heparin also contains 3-*O*-sulfate groups which don't bind ATIII. Their functions and binding proteins still remain unknown. Co-crystallization experiments suggested that different HS3ST isozymes may have distinct requirements of substrates, which are products undergoing tightly controlled sulfation (2-*N*, 6-*O* on glucosamine residue, and 2-*O* on hexuronic acid) and epimerization. Studies on the expression of *hs3st* genes further showed that they are controlled in a spatial and temporal manner. Although a HS motif does not necessarily require 3-*O*-sulfation, it is likely that a special 3-*O*-sulfation represents the control of a special biological function. It will be highly interesting to look for 3-*O*-sulfation containing HS oligosaccharides and identify HSBPs that bind specifically to them.

As mentioned before, multiple factors contribute to the binding affinity between HS and HSBP. As a result, the binding may depend on an individual sulfate group, a small subset of sulfation/epimerization arrangements, or simply the sulfation degree. Given the paradoxical information regarding the HS binding mechanisms, studies supporting any of the scenarios cannot be satisfactory without illustrating the exact binding sequences on HS chain. On the other hand, since a HS sequence is capable of binding to multiple proteins (albeit with different affinities and specificities), study of HS fine structure will contribute to designing tailored HS oligosaccharides with high selectivity. A successful example is that Fondaparinux (trade name Arixtra) significantly reduces the risk of heparin-induced thrombocytopenia, which is a severe side effect of unfractionated heparin.

To date, multiple techniques have been used in determining the HS sequence and its binding with HSBPs, such as enzyme digestion followed by polyacrylamide gel electrophoresis (PAGE) or liquid chromatography (LC), mass spectrometry (MS), multi-dimensional nuclear magnetic resonance (NMR) spectroscopy, and X-Ray crystallography (20). Among these methods, mass spectrometry provides promising results in high-throughput omics study and rapidly becomes the most important tool in glycomics (including GAG) analysis (21). By coupling with chromatography methods, MS is able to detect low abundant products from digestion mixtures. With improved fragmentation methods on high resolution tandem mass spectrometry, researchers can acquire information about fragment compositions (including sulfur number), linkage types, and probably epimer (22).

Chapter 2 Review of Sequencing Methods in Proteomics and Glycomics Using Tandem Mass Spectrometry

HS identification has much in common with the identification of other biopolymers (peptides, glycans and glycopeptides): all these molecules are formed as sequences of defined residue units, and their fragments in the tandem mass spectra reflect subsets of the precursor ion compositions. Therefore, identification algorithms for one type of molecules are in theory portable to other types. Due to the paucity of HS sequencing algorithms (20), it is important to recognize the issues that have been addressed before in related fields so that the corresponding strategies can be migrated into HS identification. Issues that are still open or peculiar to HS analysis should be discussed separately with caution.

2.1 Definition of Sequencing Problem

In a typical shotgun proteomics study, protein sequences are selectively digested into peptides during proteolysis. Trypsin is the most commonly used enzyme to cleave the sequences at the carboxyl side of arginine and lysine, with the exception that either is followed by proline. The peptide mixture undergoes separation through liquid chromatography (LC) and ionization in a mass spectrometer ion source. Designated peptide ions (precursor ions) are further dissociated into fragments (product ions) through multiple fragmentation pathways, and the product ions' *m/z* and intensity information is recorded in the tandem mass spectra. Peptide sequencing, in general, is about finding a definitive sequence that connects the product ions to the belonging precursor ion. The

most common way is to formalize the procedure using prior knowledge of the sequences: selecting the optimal sequence(s) from a set of candidates within the error window so that the predicted fragments from the selection should match the actual spectra closely.

Collision-based dissociation is widely used for biomolecules, including, collision-induced dissociation (CID) (also known as collisionally activated dissociation (CAD)) and high energy collision dissociation (HCD). Activated electron dissociation (ExD) methods commonly used in proteomics and glycomics include electron-capture dissociation (ECD), electron-transfer dissociation (ETD), and electron detachment dissociation (EDD). These methods provide distinct fragmentation patterns and preferred fragment types, which can be used to assist the identification. For proteomics database search methods, the search space is limited to peptide sequences derived from protein databases, or genomic database after six-frame translation. For *de novo* sequencing (23) where no prior sequence information is assumed, the search space expands to all possible peptide sequences. In this context, *de novo* sequencing is viewed as the generalization of database search methods, and only considered when the species is unknown, the database is unavailable, or as an orthogonal method to database searching.

Both database search and searching-based *de novo* sequencing methods make several assumptions implicitly. The first assumption requires that the actual sequences or at least the homologous sequences have to be available for searching. This means the best candidate sequences are selected through comparison over sequence pool. The second assumption states that the matched peaks should in general refer to preferred product ions based on instrument-specific fragmentation rule and chemical preference.

Therefore, if a peak matches to a B_2 ion, then it is scored as a B_2 ion. Once a second product ion also matches to the peak, different algorithms may choose to ignore or overcount the second ion. The first assumption may not hold if the sample comes from unknown species, or there is no available database. Alternative splicing also causes trouble in identification. The second assumption may fail if the product ion types are dubious. This is frequently observed in the presence of internal fragments, or in phosphopeptide identification where the phosphate group is labile.

Peptide identification can also be solved without exploring the sequence space. Peaks with informative mass differences in the tandem mass spectra can be assigned to residues. Connected residues eventually form one or more paths that represent the candidate sequences. This intuitive way of sequencing is widely used in manual interpretation of tandem mass spectra (24). In cases where the full sequence is hard to acquire, subsequences with a small number of residues (sequence tags) are generated and coupled to database search methods for further identification (25). For *de novo* sequencing, the identification problem is considered as finding the longest path in a spectrum graph (26).

The identification of the whole protein sequences requires the collection of identified peptide sequences, and mapping from peptides to proteins in the database. Different proteins may share identical peptides and some proteins may contain one-hit peptides. Models have been proposed to infer the most likely set of proteins. It is worth noting that the protein vs. peptide relationship is analogous to the fragment ion vs. ion mass relationship, both of which can be illustrated as a bipartite graph. Protein inference

is a complicate problem involving uncertain peptide identification, degenerate peptides, peptide detectability and other issues (27), which is beyond the scope of the discussion in the thesis.

If the main goal is to study the modification (*e.g.* phosphorylation) localization along the sequence instead of the sequence itself, the sequence identification problem can be framed as distributing a set of pre-determined modifications to possible candidate residues, which is essentially a combinatorics problem. Peptides with modification may compete with intact peptides for peptide-spectrum match (PSM) by generating a large number of dubious fragments. Modified peptides may also affect the specificity of enzyme cleavage (28). Besides, the modification may be lost during fragmentation. These all adds extra difficulties in recovering the sequence structure.

Glycan sequencing entails building a glycan tree where each node represents a residue and the edge is the linkage between residues (29). The sequencing methods share basic procedures of peptide sequencing. Candidate oligosaccharides can either be selected through database searching (30), or constructed in a *de novo* way (31).

Glycopeptide sequencing is more than the combination of identifying peptide and oligosaccharide sequences. The glycosylation site specificity is important for illustrating the biological functions of glycopeptide, and currently the focus of glycopeptide characterization. Due to the lack of standards and coisolation of proteoforms, the sequencing problem may also be framed as sequencing from mixtures, which has been discussed before (32).

2.2 Review of Database Search Methods

The earliest database search methods date back to 20 years ago. In 1994, Mann and Wilm introduced PeptideSearch (25) to extract sequence tags from tandem mass spectra and search them against the sequence database. In the same year, Eng *et al.* (33) implemented SEQUEST, an automated database search method based on cross-correlation function. Since then, several other database search methods have been developed, including Mascot (34), X!Tandem(35), OMSSA (36), MyriMatch (37), ProteinProspector (38, 39), Andromeda (40), Morpheus (41), Comet (42), and MS Amanda (43). Database search methods are widely used in peptide and protein identification (44, 45), and remain part of the standard procedure for modern proteomics analysis workflow (46, 47). Despite distinct features provided by different search engines, most methods follow the general procedure of matching an experimental tandem mass spectrum against a genomic database. During the matching procedure, the quality and organization forms of the database, the characteristics of the experimental spectra, as well as the scoring function measuring the fitness between the spectrum and entries in the database may all affect the final identification performance.

A spectrum is simply a series of *m/z* and intensity pairs, while a peptide sequence consists of ordered amino acid residues. In order to bridge the gap between these two forms, format transformation is necessary. Researchers can either choose to match the observed spectra against theoretical spectra, which are generated from the protein sequences in the database via *in silico* digestion and fragmentation, or against annotated spectra from previous experiments. Alternatively, one can convert the target spectra into

sequence or sequence tags (*de novo* sequencing), and then searched the sequence or tags against the protein sequence database (48). Most database search tools support sequence database in the format of FASTA (49). The sequence files can be downloaded from databases targeting specific organisms, or comprehensive non-redundant databases containing many taxa, e.g. UniProt (50) and RefSeq (51), or large redundant database such as Entrez Protein. When the protein sequences for the studied species are unavailable, it is also possible to search against nucleic acid databases, such as DNA, mRNA and expressed sequence tags (ESTs) after six-frame translation (52).

Trypsin is the most widely used protease in shotgun proteomics for protein digestion. It typically cleaves protein sequences at the carboxyl side of lysine and arginine, unless the target residue is followed by a proline. Alternative enzymes, such as Lys-C/Chymotrypsin, are also combined with trypsin in experiments in order to improve the coverage of whole protein sequencing. Enzyme specificity (full, semi or non-specific) and missing cleavages (0, 1, or 2) may are important factors determining the PSM, and researchers are responsible for configuring the parameters based on the goal of the experiments.

Pioneer database search methods originated with the availability of low resolution tandem mass spectrometer (*e.g.* linear ion trap). The data produced contain a large amount of peaks with missing isotopic patterns. As a result, charge state information for peaks may be lost and use of a probabilistic model is needed for estimating instrument-specific ion types, charge states, and likelihood of spurious peaks. As high accuracy and high resolution instruments such as QTOF, and FTMS (including

ion cyclotron resonance and Orbitrap) emerged, the isotopic clusters for fragment ions are more likely to be maintained. Mass differences between isobaric ions due to isotope fine structure are also able to be detected for low masses. With the widespread availability of such high quality data, algorithms including Morpheus (41) and MS Amanda (43), were developed to use non-probabilistic scoring functions to produce faster and better identification results relative to classical methods.

Another factor that affects the identification of a sequence is the scoring function, which serves to differentiate the correct sequence from incorrect sequences. Sadygov *et al.* (44) categorized the PSM scoring function into four types: descriptive, interpretative, stochastic and probability-based. The scoring function can be based on different probability models, such as binomial distribution for Andromeda (40), or can be simple counting strategy, as implemented by Morpheus (41) .

Post-translational modifications (PTMs) expand the functional diversity of linear protein sequences, and increase the number of structural variants exponentially. For database search methods, the existence of PTMs poses dual effects on the actual identification. On one hand, they serve to double check the identification of native sequences. Some search tools require the native sequences to be identified in the first run and modified versions in future runs, which is a strategy adopted in multiple-pass searches. On the other hand, PTMs on sequences may be misidentified as different native sequences, requiring researchers to manually check the data. False positives are difficult to detect unless the modification type is explicitly specified, or an unrestricted PTM

identification method is adopted (53, 54). Most database search tools allow users to specify fixed (static) and flexible modifications.

A fixed modification, such as phosphorylation, has the same effect as updating the mass of the modified residue, which causes no extra burden in searching. The presence of two types of modification at a given site doubles the search time for each candidate modification site, and the combination of multiple modification types exponentially increase the search time. Therefore, strategies have to be made to control the complexity of searching conditions. Andromeda (40), the search engine of the quantification software MaxQuant, exhaustively lists all the possible combinations of PTMs on the protein sequence to improve the identification rates. It uses multiple levels of indexing to allow the program to run smoothly on a laptop. If glycosylation is taken into account as a peptide PTM, the total searching space will further increase by orders of magnitude, making either the standard database search or manual specification of sequences unfavorable. Nearly all mainstream database search tools consider only the number of modifications on the sequence during MS1 candidate filtering. Such tools do not perform well for locating the exact modification sites (modification localization). Usually dedicated tools are needed if the goal of experiments is to identify modification localization.

Due to the complexity of factors determining the match performance, it is usually difficult for novices to develop an appropriate analysis workflow; however, guidelines on analyzing the data and discussion of potential problems are available (55). Nonetheless, it is often necessary for users to manually verify results. This seems acceptable in a high-

throughput study, since the parameters only need to be tuned once. However, it raises the difficulty of combining datasets across multiple experiments, especially those from different labs. Machine learning techniques were considered to assist the automatic optimization of parameters, but in most cases, despite steady progress in proteomics tool suites, proteomics data analysis remains to a large degree a time consuming effort.

Few standard datasets are publicly available allowing developers and users to test their search methods (56). In most cases, the target-decoy approach (TDA) (57) is used to measure the specificity by assuming that decoy sequences are randomly distributed among the search space, which provides a good measure of false discovery rate (FDR). The TDA method provides a neat and accurate way of validating proteomics results, and threshold of 1% or 5% are widely used by almost all database search tools. However, cautions should be taken on the validation report when using a TDA method (58). Use of too small a searching space, biased sequence distribution due to multi-pass search, errors in translation from nucleic acid database, and among other problems, may skew the background distribution and produce unexpected results on the normal FDR threshold using TDA approach.

The TDA approach facilitates the peptide identification in that once the FDR threshold is well controlled, the more peptides a search tool identifies, the better performance it provides. This allows the strategy of combining searching results from multiple engines to improve the overall identification rate. The basic assumption of this strategy is that different tools may excel in identifying certain types of peptides for certain instruments, but deteriorates in other types of peptides. A set of tools were

developed to support the combining method, such as MSBlender (59), PepArML(60), ConsensusID (61) module of OpenMS (62), iProphet (63), and PeptideShaker (64, 65). Shteynberg *et al.* (66) reviewed the identification results produced by different ways of combining search engines, and suggested to learn the complementarity and similarity between tools to avoid unnecessary time waste and bias towards certain groups of peptides. More strategies have also been proposed to improve the identification rates or reducing false positives, which includes combination of multiple fragmentation modes (67) or other omics data, digestion with multiple enzymes, multi-pass search, improvement of validation approach, construction of customized database from genomic database or RNASeq data.

As discussed above, informatics workflows for proteomics is still in immature stage. Considerable user experience and efforts are necessary to monitor the quality of the analysis. The problem may root in the design of scoring functions, which favor a set of peaks while disregarding the rest. There are chances that the favored peaks may have been mistakenly assigned or weighted due to unknown dissociation mechanisms, internal cleavage, or PTM. TDA does not provide a safety net to prevent the search tools from making such mistakes. Further, the uncertainty from peptide identification will amplify when used to infer protein structure, which makes the inferred results even less reliable.

2.3 Review of *De Novo* Sequencing Methods

Due to the rigorous assumption of sequence availability, *de novo* sequencing is considered as a complementary method to validate identification results from database

searches. It is also extremely useful when the matching of a homologous sequence database is unavailable (*e.g.* snake venom antibodies), the organism in question mutates rapidly (*e.g.* some viral pathogens) or when species or organism is unknown. Current *de novo* sequencing methods can be categorized into naïve approaches, spectrum graph models, probabilistic and combinatorics models (23) depending on their specific assumptions. The definition of *de novo* sequencing is very rough, since the naïve approach, which enumerates all or set of the theoretical possible sequences and finds the optimal one, is similar to a database search method. Therefore, many researchers simply view *de novo* sequencing as a general case of database search (23, 68). As a result, the PSM scoring function, which typically requires assumption of the fragmentation patterns in the tandem mass spectra, can be migrated seamlessly from database search into *de novo* sequencing.

Naïve approaches simulate the method of database searching, while adopting different strategies to reduce the searching space. PEAKS (69), which is the most popular commercial software for *de novo* sequencing, generates 10^5 sequences and merges the sequences into a consensus sequence with local confidence on residues. Heredia-Langner *et al.* (70) implemented a genetic algorithm (GA) to optimize the candidate sequence efficiently from a very large search space. An advantage of naïve approaches is that they tolerate the missing fragments and internal fragments that are often difficult to address by other *de novo* sequencing algorithm, but they may be less attractive when the protein sequence database is accessible.

Another major branch of *de novo* sequencing consists of graph-based models. These approaches attempt to represent the tandem mass spectrum as a graph with the vertex representing observed peak and the edge representing amino acid residue mapped from the mass difference between vertexes. The optimal sequence corresponds to the longest path in the graph, and different methods were proposed to solve the problem based on dynamic programming techniques (26, 71, 69, 72). Due to the ambiguity and unknown fragmentation mechanism, even the optimal sequence may not be correct. Therefore, algorithms looking for suboptimal solutions were also developed (73). The computation time usually is a function of the number of nodes multiplied by the number of edges, which might be huge if all the peaks are considered. As with database search methods, emergence of high-resolution tandem MS data also contributes to the identification quality and proteomics-grade sequencing results (74).

There are a few algorithms that utilize the features of tandem mass spectra and peptides. Zhang *et al.* (75) designed a divide-and-conquer method to separate the whole spectrum into subspectra, upon which the sequencing task is running. Spengler (76) proposed a composition-based sequencing strategy to sequence peptides based on peptide composition and high accuracy mass spectrometer. Olson *et al.* (77) provided an improved composition-based sequencing algorithm working for MALDI TOF/TOF. It contains an amino acid composition look-up table, and retrieves the amino acid compositions from the b- and y-ions. By expanding the compositions of the ions, from single amino acid till the full peptide composition, the actual sequence can be decided.

Common complaints regarding *de novo* sequencing methods include time-consuming performance, poor identification rate and coverage, and lack of validation methods. For spectrum graph methods, the missing fragments and ambiguous paths may prevent an algorithm from finding the optimal sequence. For naïve methods, the enumeration of search space undergoes combinatorial explosion, which maybe intractable as the sequence length grows. Significant efforts have been made to improve the covered sequence length and accuracy, which include: combination of spectra pairs from multiple fragmentation mode, *e.g.* collision-induced dissociation (CID), electron-transfer dissociation (ETD), and high-energy collision dissociation (HCD), and multiple enzymes (78); combination of spectra from top-down and bottom-up proteomics experiments (79); and appending *de novo* sequencing with homologous database search (48). The emergence of high-resolution MS/MS data also contributes to the identification quality and proteomics-grade sequencing results (74).

2.4 Identification Methods in Post-translational Modification Study

Post-translational modifications modify protein sequences, adhesive properties and are required for all aspects of physiology. There have been hundreds of modification types reported, and stored in databases, including Unimod (80) and RESID (81). The identification of PTMs on peptide sequences is a non-trivial computational problem. The complexity of PTM analysis lies in the fact that PTM triggers new varieties on the sequences as well as the spectra, where traditional identification methods were not designed for addressing those issues. Na and Paek (68) summarized the issues into

several categories. One of the problems is that the dynamic range across modified and unmodified peptides is terrifically large for complex samples, *e.g.* from 1 to 10^{11} for human plasma proteome (82). Modified peptides with low abundances may not be identified. The second problem is that some types of modification are labile and significant peaks with neutral losses are observed, for example phosphorylation using CID. The third problem is that peptides with different site-specific modifications may co-elute and co-fragment, a problem that is very common for histone proteins (83). The modification can also exist in multiple forms, *e.g.* glycosylation, which may dissociate during fragmentation procedure. Finally, the existence of a modification may significantly change the fragmentation pattern of the peptide, while produce specific diagnostic ions, such as immonium ions (28). In early days, most PTM identification studies were performed through database search methods by configuring static and flexible modification types upon the restricted amino acid residues. This quickly becomes computationally intractable when a large number of modification types are taken into account.

Compared with an unmodified one, a modified peptide shifts part of the generated product ions towards increasing *m/z*. In the case of only one modification on the peptide, the starting position of the shifts usually indicates the modification site and the shifted distance represents the modified chemical group. In other words, the modification contributes to dividing the spectrum into two separate sub-spectra, for each of which the relative *m/z* positions of the peaks are well maintained.

In order to reduce the computational cost during exhaustive modification type searching, most modern PTM algorithms adopt multi-pass strategies to remove unrelated candidate peptides. In the first pass, peptides with the minimum number of modifications (usually due to sample preparation) are searched and identified. In the second pass, peptides identified from last round are combined with a comprehensive list of known modifications and re-searched to explore possible types of modifications. The second step is usually called blind search, or unrestrictive search, comparing to "restrictive" search in classical database search methods that target for peptides with minimum modifications. Blind searches compare the tandem mass spectra against candidate peptides for all known or even novel types of PTMs. As more types of PTM are considered, the search space expands exponentially. However, the size of the search space remains acceptable if the total number of candidate peptides is well controlled.

Different methods have been proposed to improve the matching performance between the tandem mass spectra and modified peptides. MS-Alignment (84, 85), the first blind search algorithm, finds optimal alignment between spectrum and peptide using dynamic programming method. OpenSea (86) and SPIDER (87) look for the difference between *de novo* sequencing and unmodified peptide candidates. MODa (88) provides a fast solution to identify peptides with multiple unknown modifications based on spectral alignment (85) and sequence tag. As discussed above, the sub-spectra separated by the modification maintain the relative *m/z* positions of the bounded peaks. The relationship between spectra from unmodified and modified peptides has also been studied. The notion "spectral pair" notion is used to represent the closely related spectra. ModifiComb

(89) and DeltAMT (90) groups spectral pair with difference precursor mass values but similar retention time and predict putative modifications. Spectral network (91, 92) organizes the spectra from the network perspective, where each node represents a spectrum and each edge represents the similarity relationship between nodes. As a result, spectra similar to each other (even partially) are clustered together and the knowledge of unknown spectra can be inferred from spectra with annotations.

As more types of modifications are taken into account, the trade-off between sensitivity and specificity should also be given attention. By adding an additional modification mass, the spectrum can be better explained. However, effect might just be caused by spurious peak annotations instead of a novel modification. As a result, a much higher scoring threshold is needed to distinguish the identified modified peptide sequences from the background. Another issue is the "skewed" distribution of the scores from decoy sequences due to the small set of candidate peptides. Efforts have also been made on designing robust scoring function and reconstructing decoy database.

Most PTM tools focus on the identification of types of modifications, while ignoring the modification localization issue posed by PTM. Different from the identification of modification types, where the mass values of the product ions and precursor ions all contribute to the identification, only a few peaks (one b-ion and y-ion in the case of one modification) are able to suggest the actual modification site. Missing fragmentation, low-abundance peaks, and spurious peak interpretation may mislead the identification of modification localization. AScore (93) evaluates the confidence of phosphorylation sites based on site-determining ions, by considering the phosphorylation

on each candidate site. SLoMo (94) extends the AScore algorithm to cases with multiple instruments, multiple search engine and different modification types in Unimod (80). The presence of isomers may make the problem more challenging. A few methods have been proposed to address the issue (83, 95).

2.5 Identification methods in Glycoproteomics

Glycans, or carbohydrates, are critical in regulating cell signaling, cell-cell adhesion, and pathogen recognition. It has been estimated that at least half of proteins in eukaryotes are glycosylated (96), although the latest study suggested the number be one fifth (97). In addition to the peptide sequence, a typical glycoproteomics study also covers the identification of compositions of carbohydrate moieties and site-specific connections between the carbohydrate and peptide sequence.

The identification method for glycoproteomics is still in its infancy stage. Efforts have been focused on the development of software suite or home-made scripts to accommodate specific experimental design, but little generality has been achieved to facilitate the community (98–100). One of the possible reasons is that new sample preparation methods are emerging to achieve better sensitivity and higher throughput, which complicates the data processing and require dedicated tool. Another cause may be the lack of consensus in defining milestones in glycoproteomics pipeline so that software project from one group cannot be easily followed and improved by another group. Besides, PTM identification problem in proteomics is still open and the addition of

microheterogeneity and macroheterogeneity from carbohydrate moieties complicates the task.

2.6 Challenges in Heparan Sulfate Sequencing

Identification of HS molecules shares some similarities with peptide identification. All glycan structures have clearly defined set of residues and fragment types, which was first proposed by Domon and Costello (101). Therefore, routine strategies in proteomics can be considered for HS sequencing. From the perspective of database searching, one can simply construct a HS sequence database based on known biosynthetic rules (102), or enumerate all theoretically possible sequences. The scoring function and FDR calculation in proteomics may assist the algorithm development for HS sequencing. Potential problems may arise for a database version of HS sequencing, which include: the sequence space undergoes combinatorial growth as the chain length increases; top-ranking sequences may not differentiate with each other, even for pure standards; biosynthetic rules summarized from previous studies are not comprehensive and rare sulfation pattern may associate with special biological function. A traditional bottom-up *de novo* sequencing strategy is still problematic, as the mass difference between two assigned peaks is very likely to match to residues with sulfate loss, or more than one possible answer. Besides, sulfate groups on the HS sequence is comparable to PTM on peptide sequence, where both have preferred candidate modification sites and modification events will shift the related peaks by a fixed mass value. Methods for modification localization are potentially valuable for identifying the sulfation localization on HS chain. There were several pioneer studies on automated HS sequencing, as

discussed below. However, compared with the fast evolvement of instrument, the development of computational methods still lags behind.

In 2005, Saad *et. al*. developed HOST, a spreadsheet-based heparin sequencing tool using enzymatic digestion and a CID ESI-MS^n approach (103). HOST built oligosaccharide sequences based on similarity analysis that compares tandem mass spectra of disaccharide units produced from the saccharide in question against those acquired from the intact saccharide. Methods have advanced since that time and now allow determination of sulfation and acetylation directly from tandem mass spectra without separate enzymatic digestion. Simulation model (102) was also explored to predict the fine structure and domain organization of HS sequence using information from enzymatic digestion and Golgi-based biosynthetic rules. This method produced an "average" HS chain statistically and is valuable for guiding the selection of candidate sequences, yet it failed to pinpoint the positions of sulfate groups for a specific chain. The public tool Glycoworkbench (104, 105) was also developed to facilitate the assignment of monoisotopic peaks in tandem mass spectra of glycans (including GAG), but it aided little in the identification of sulfated sites on the sequence scale.

Chapter 3 *De Novo* Sequencing of Heparan Sulfate Using Synthetic Pure Standards

3.1 Introduction

The identification of HS fine structure has long been a labor-intensive and error-prone process, which could be attributed to the lability of sulfate groups during fragmentation as well as ambiguous structural interpretations associated with each identified monoisotopic peak. As discussed in Chapter 2, the traditional sequence identification methods are not applicable to HS sequencing: there is no available public database for HS sequence, and the total number of candidate isomers increase combinatorially, depending on the candidate sulfation (and acetylation) sites as well as the number of sulfate groups; one candidate isomer might be very similar to another in terms of their theoretical spectra. On the other hand, the mass difference of two peaks can easily match to monosaccharide residues plus certain number of sulfate groups, which may simply result from random match.

A method for automatic HS sequencing will not succeed without considering the fragmentation pattern. However, the pattern summarized based on human expertise may easily be distorted by alternative explanations, especially for structure consisting of identical components. In order to solve the problem, I designed HS-SEQ (106), the first HS sequencing algorithm using high resolution tandem mass spectrometry. It focuses on high-confidence peak interpretations instead of global sequence-spectrum match. This chapter will introduce the principle and framework of HS-SEQ in solving the HS sequencing problem.

3.2 Summary of Technique Advance for Heparan Sulfate Fragmentation

The development of experimental methods over the past few years has turned the prospect of HS sequencing into reality (107–109). Recent breakthroughs in chemoenzymatic synthesis of HS with specified sulfation positions have removed roadblocks in designing substrates with desired properties (110). Meanwhile, new tandem mass spectrometric dissociation techniques are now capable of identifying sulfation patterns of some GAG oligosaccharides directly (17, 18). In particular, electron detachment dissociation (EDD) (22)and negative electron transfer dissociation (NETD)(111, 112) techniques generate informative spectra that include rich glycosidic-bond and cross-ring cleavages while maintaining the intact sulfation information. It is also possible to suppress losses of sulfate groups in collision-induced dissociation (CID) tandem mass spectra by derivatizing HS saccharides (113) or by replacing acidic protons with metal cations (114, 115). However, all tandem mass spectra of highly sulfated HS saccharides display useful backbone dissociation combined with some degree of sulfate losses. This gives rise to a multiplicity of product ions with either intact or reduced sulfate numbers, the interpretation of which limits the dissemination of these techniques. It is thus necessary to develop commensurate computational methods to meet the challenges of these new techniques, which will promote the systematic study of the *in vitro* and *in vivo* effects of HS fine structure on physiological activities.

3.3 Definition of Heparan Sulfate Sequencing Problem

For HS, the size of the sequence space is a combinatorial function that depends on the oligosaccharide backbone and numbers of acetate/sulfate groups. For example, a

hexasaccharide with composition [1, 2, 3, 1, 6] ([ΔHexA, HexA, GlcN, Ac, SO_3]) gives rise to 1,386 isomers, while a 14-mer oligosaccharide with composition [0, 7, 7, 2, 5] produces 1,381,380 theoretical sequences. Sampling methods may speed up the searching process, but the resulting local maximum and convergence problems pose additional burdens on configuring reasonable searching/scoring schemes. Even with efficient limitation of the search space, it may still be difficult to distinguish the true sequence from candidate sequences via sequence scores, since the regions with incorrect sulfate/acetate numbers may be supported by falsely assigned product ions due to product ion assignment ambiguity.

The HS sequencing problem can be considered as finding the best way to distribute fixed numbers of acetate/sulfate groups along the precursor sequence. The precursor mass usually uniquely determines the HS composition, *i.e.* the counts of monosaccharide residues, sulfate and acetate groups, whereas the counts of monosaccharide residues implicitly determines the sequence of monosaccharide residues and therefore the sequence of candidate acetylation/sulfation sites along the precursor sequence. In this sense, we can rephrase the HS sequence problem as finding the best way to distribute a fixed number of acetate/sulfate groups among a fixed set of candidate acetylation/sulfation sites.

We clarify several terms (Figure 4B) here in order to facilitate the discussion. We use modification to represent acetylation/sulfation, the locations of which are uncertain on the precursor sequence before prediction. In contrast, we consider derivatization at the reducing-end as an integrated part of the precursor backbone since there is no

uncertainty of its location. Accordingly, a candidate modification site represents a particular position on a particular monosaccharide residue where the modification may occur, and modification number stands for the count of the modification occurrence. The solution is a list of modification distributions that specify how the given numbers of modifications are distributed among all the candidate modification sites. Each modification type corresponds to a modification distribution, and the sum of the distribution equals the modification number defined in the precursor composition.

The concepts of candidate modification sites and modification number serve as the building blocks for the whole framework of HS-SEQ. For a given peak, we denote each structural interpretation of a peak (the ion type, cleavage position, neutral loss and modification numbers, e.g. $Y_3 - H_2O + 1Ac + 3SO_3$) as an assignment. Each assignment in essence contributes structural information regarding the covered candidate modification sites and modification numbers (Figure 4C). In this sense, each assignment defines a subproblem of the original HS sequencing problem, which serves to update the global modification distributions to finer resolution (Figure 4D). Moreover, a terminal assignment describes an assignment containing an intact monosaccharide residue of either the reducing end (RE) or non-reducing end (NRE), and an internal assignment describes an assignment containing no intact RE or NRE residue. For each modification type, we use assignment graph to describe the relationship between assignments. In the graph, the node represents an assignment and the edge represents the inclusion relationship of candidate modification sites of two assignments.

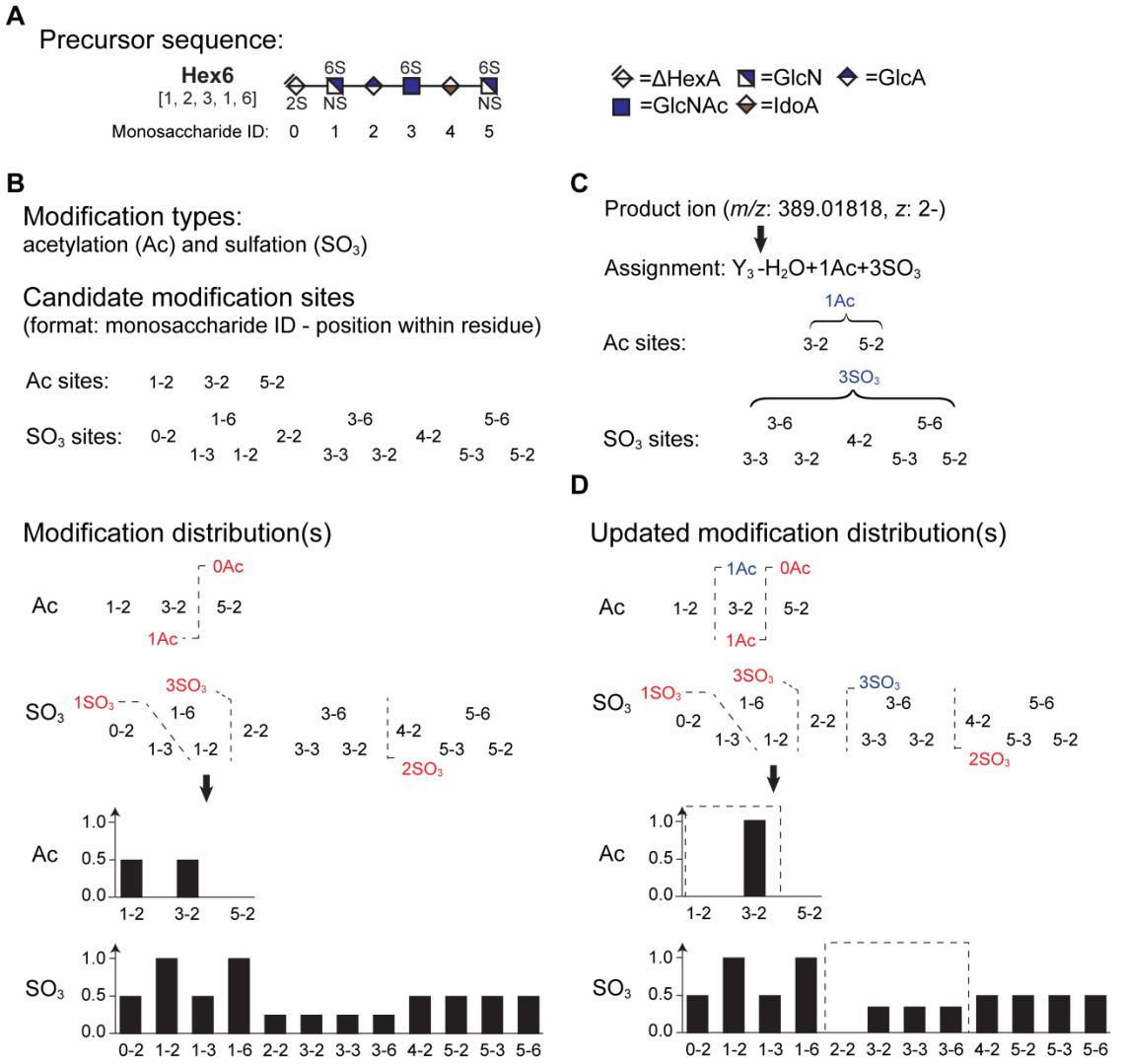
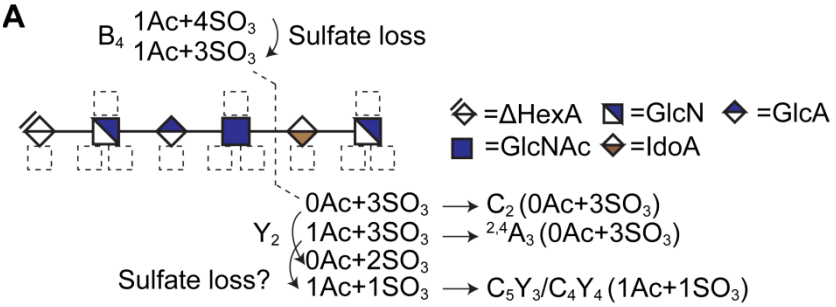


Figure 4 Summary of basic concepts in HS-SEQ. (A) The hexasaccharide sequence [1, 2, 3, 1, 6] ([ΔHexA, HexA, GlcN, Ac, SO_3]) illustrated by cartoon symbols. (B) Visual representation of the hexasaccharide sequence in HS-SEQ. (C) Visual representation of assignment in HS-SEQ. (D) Assignment updates the modification distribution. The original modification numbers are denoted by red text and the new numbers deduced from the assignment are denoted by blue text. The updated regions are marked by dashed squares.

In order to identify the HS sequence, the intuitive way is to map the mass values of peaks to assignments (spectra interpretation), and collect the structural information of assignments to deduce the modification distributions of the precursor sequence (sequence assembly). Ambiguity occurs when a single mass value corresponds to multiple assignments, and different assignments produce inconsistent structural information regarding the precursor sequence (Figure 5A). In essence, the data ambiguity problem is all about confusing information of the candidate modification sites and/or modification numbers (Figure 5B). As we have seen, nearly all the concepts and difficulties in HS sequencing can be rephrased in terms of candidate modification sites and modification number. HS-SEQ relies on the rephrased concepts and framework built upon to address the HS sequencing problem.

A



B

	Modification sites	Modification numbers	Example
Identical mass	S	D	Y_2(1Ac+3SO_3) and $^{2,4}A_3$(0Ac+3SO_3)
	D	S	Y_2(1Ac+1SO_3) and C_5Y_3(1Ac+1SO_3)
	D	D	Y_2(1Ac+3SO_3) and $^{2,4}A_4Y_5$(0Ac+3SO_3)
Different masses	S	D	B_4(1Ac+4SO_3) and B_4(1Ac+3SO_3), B_3(0Ac+3SO_3) and $^{1,5}A_3$(1Ac+2SO_3)

Figure 5 Data ambiguity in HS sequencing. (A) Assignments of B_4 and Y_2 on a hexsaccharide with composition [1, 2, 3, 1, 6]. Some Y_2 assignments have incompatible

modification numbers with the B_4 assignments. For example, B_4 ($1Ac+4SO_3$) and Y_2 ($1Ac+3SO_3$) cannot co-exist since the total number of Ac on the sequence is only 1. Alternative assignments are suggested after the arrows. For example, Y_2 ($0Ac+2SO_3$) can be considered as Y_2 ($0Ac+3SO_3$) with sulfate loss, or C_2($0Ac+3SO_3$) with sulfate loss. (B) Classes of data ambiguity. Assignments with either the same mass values (isomeric or isobaric) or different mass values can cause ambiguity, and the ambiguity in essence is the ambiguity of the candidate modification sites and/or associated modification number. "S" denotes "same" and "D" means "different".

As shown in Figure 6A, the sequencing task in HS-SEQ consisted of two steps: 1) the prediction of the distribution of acetate groups, taking into account the data ambiguity; and 2) the prediction of the distributions of sulfate groups, a step divided into sulfate numbers on each monosaccharide residue and exact sulfation positions within residues. Sulfate loss was considered during this step. Figure 6B illustrates how an assignment connected to others and contributed to updating the modification distribution. HS-SEQ organized assignments by their respective confidence values and sequentially inserted the assignments into the assignment graph. The insertion of each assignment further updated the modification distribution. The final modification distribution can be read manually, or converted to a list of top candidate sequences (see discussion below).

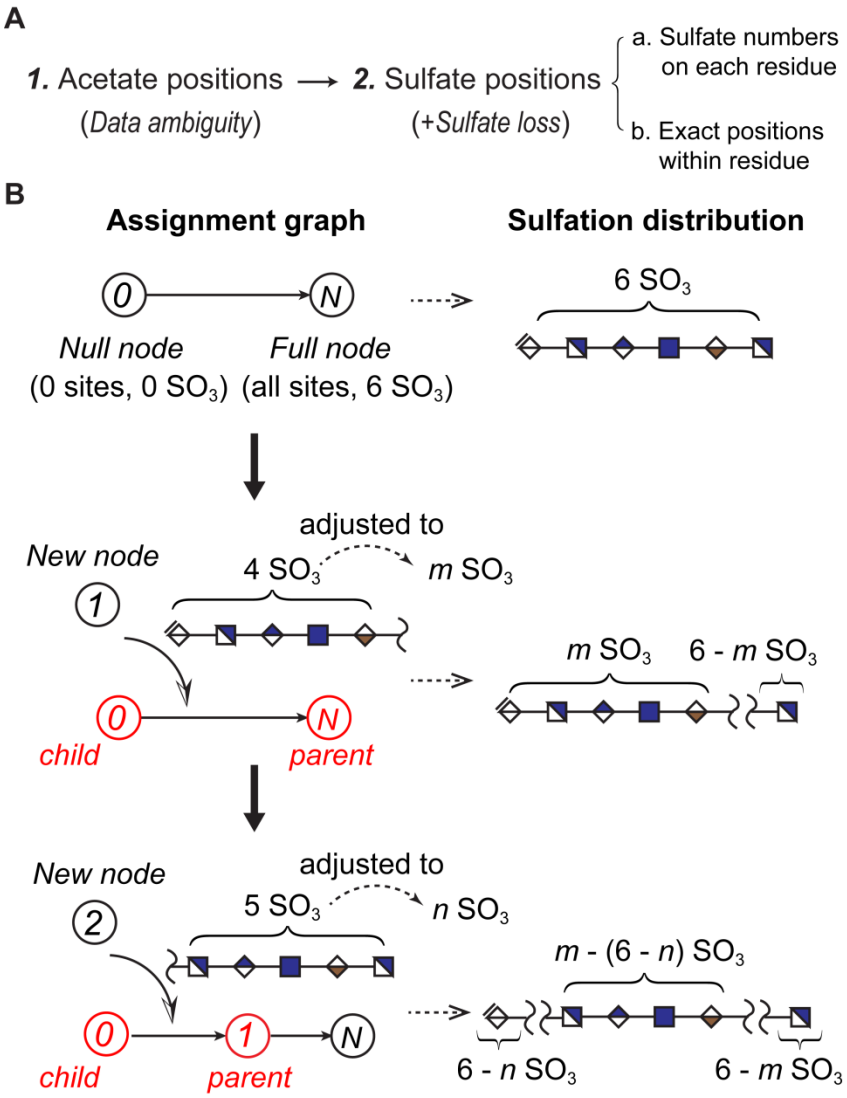


Figure 6 Schema of HS sequencing in HS-SEQ. (A) Subtasks in HS sequencing. In HS-SEQ, HS sequencing consists of two basic steps: identification of Ac positions and identification of sulfate positions. Data ambiguity is considered for each step, and sulfate loss is considered when identifying sulfate positions. (B) Assignment graph connects assignments and generates the modification distribution.

We tested HS-SEQ using 9 synthetic HS saccharide standards representing a range of chain lengths, modification positions, sulfation degrees, reducing-end derivatization groups, and ion charge states. The results showed that HS-SEQ accurately recovered the correct HS sequences for 76% (19 out of 25) of the tested tandem mass spectra, and approached the correct sequences for the remainder. For each oligosaccharide sequence, at least 50% of the tandem spectra (each with a different charge state) reported the true sequence as rank 1. Moreover, the scores for the correct HS structures were distinct from their isomeric candidates, and the computation time required for sequencing was usually a few seconds, demonstrating the feasibility of a HS high-throughput sequencing pipeline. The program was developed using the C++ language and is available for download through http://code.google.com/p/glycan-pipeline/. The program currently runs under command line and requires only a few basic arguments (precursor ion *m/z*, charge state and input/output file). It also includes XML configuration files that specify the rest of parameters and their default values. A graphical-user-interface (GUI) version of the program is under development and will be available in future version.

HS-SEQ is the first tool to systematically study the problem of automatic HS sequencing. It is based on a unique sequencing strategy to takes advantage of the latest high-accuracy and high-resolution instrument. We showed HS-SEQ's capability in addressing the challenges of data redundancy, data ambiguity and sulfate loss inherent in tandem mass spectra of HS compound class, and providing production-grade distinctive results.

3.4 Method Description

3.4.1 Data acquisition and preprocessing

Tandem mass spectra of the 9 synthetic HS saccharides (Figure 7) were acquired on a 12-T solariX^{TM} hybrid Qh-Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (Bruker Daltonics, Bremen, Germany) in NETD (Figure 7). All HS standards were dissolved in 5% isopropanol, 0.2% ammonia solution to a final concentration of 5 pmol/μL, and infused directly into the mass spectrometer using an Apollo II nanoESI source. The spectra were acquired in the negative ion mode and the instrument parameters were optimized to minimize SO_3 losses. Precursor ions were isolated using a mass-filtering quadrupole and externally accumulated in a hexapole collision cell before tandem MS analysis. For NETD experiments, fluoranthene cation radicals were generated in a chemical ionization source in the presence of argon. Efficient dissociation was ensured by using a reagent accumulation time of up to 500 ms and a reaction time of up to 500 ms. Each transient was acquired with 1 M data points, each tandem mass spectrum was acquired by signal averaging up to 100 transients for improved S/N ratio. The instrument was externally calibrated using sodium-TFA clusters before tandem MS experiments.

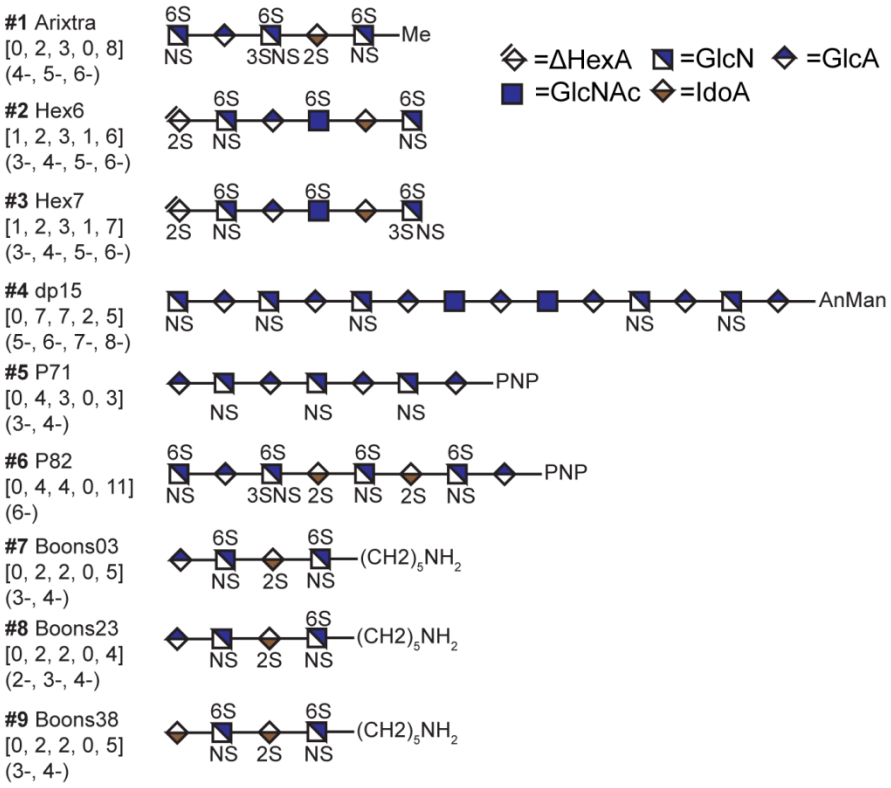


Figure 7 Structures of 9 synthetic pure standards for algorithm validation. #1 Arixtra [0, 2, 3, 0, 8] (charge state 4-, 5- and 6-) was purchased from Organon Sanofi-Synthelabo LLC (West Orange, NJ). #2 Hex6 [1, 2, 3, 1, 6] (charge state 3-, 4-, 5- and 6-) and #3 Hex7 [1, 2, 3, 1, 7] (charge state 3-, 4-, 5- and 6-) were purchased from New England BioLabs (Ipswich, MA). #4 dp15 [0, 7, 7, 2, 5] (charge state 5-, 6-, 7- and 8-), #5 P71 [0, 4, 3, 0, 3] (charge state 3- and 4-) and #6 P82 [0, 4, 4, 0, 11] (charge state 6-) were bio-enzymatically synthesized and were generously provided by Prof. Jian Liu from University of North Carolina, Chapel Hill. Synthetic HS tetrasaccharides #7 Boons03 [0, 2, 2, 0, 4] (charge state 3- and 4-), #8 Boons23[0, 2, 2, 0, 4] (charge state 2-, 3- and 4-) and #9 Boons38[0, 2, 2, 0, 5] (charge state 3- and 4-) were generously provided by Professor Geert-Jan Boons from the Complex Carbohydrate Research Center at the

University of Georgia. Me: methyl, AnMan: 2,5-anhydro-*D*-mannose, PNP: 4-Nitrophenol.

Peak lists were exported from the spectra using Bruker DataAnalysis 4.2 with the "FTMS" option selected and the signal-to-noise ratio (S/N) threshold set to 0. Each row of the peak lists contained the *m/z* value, intensity, resolving power and S/N for a given peak. An in-house deconvolution/deisotoping program (discussed in Chapter 5) was developed to convert the peak lists into lists of monoisotopic neutral masses.

3.4.2 Peak assignment

The monoisotopic peak list was converted into neutral mass values, and was matched to theoretical fragment mass with a 2 ppm mass error. The theoretical library was constructed from the composition of precursor ion [ΔHexA, HexA, GlcN, Ac, SO_3]. All possible product ions. were considered, including types of *A*, *B*, *C*, *X*, *Y*, *Z*, and internal ions generated from multiple cleavages of the HS backbone structure. Neutral mass loss (-H_2O), hydrogen transfer (-H) and all possible numbers of sulfate/acetate groups allowed on the product ions were taken into account. We assumed that for GlcN residues, sulfation may only occur at 2-*N*, 3-*O* and 6-*O* positions and acetylation at 2-*N* position, while for HexA (GlcA or IdoA) residues, sulfation may occur at the 2-*O* position. Although it has been suggested that the isomeric uronic acid residues resulted in different fragmentation patterns using EDD (22), the difference has not yet been established as an explicit pattern for automatic recognition from the spectra. Therefore, HS-SEQ did not attempt to distinguish GlcA from IdoA.

3.4.3 Sequence construction

1. *Filtering redundant data.* A typical assignment of a peak consisted of three parts: product ion type (*e.g.* B_2, C_3, and $^{1,5}A_2$), composition shift (*e.g.* $-H_2O$, -H) and modification numbers for acetate and sulfate groups. Assignments with the same product ion type and the same number of acetate groups were clustered. In other words, each cluster contained assignments with mass values differing only in the equivalent of a combination of neutral mass loss ($–H_2O$), hydrogen transfer (–H), and sulfate groups. Note that assignments of B- and C-type ions were not differentiated. Similarly, Y- and Z-type ions were clustered together. For example, assignment B_2 ($–H_2O$, 1Ac + $3SO_3$) was clustered with assignment C_2 (–H, 1Ac + $2SO_3$), but not with assignment C_2 (0Ac + $1SO_3$) due to the different acetate numbers they hold. For each cluster, the assignments with the highest number of sulfate groups were selected to represent the cluster. If more than one assignment had the highest sulfate number in one cluster, HS-SEQ chose the one by the neutral mass shift in order of priority: no shift > – H2O > +H/–H > – H_2O +H/–H. The number of +H/–H was set to be from 0 to 2 by default.

The clustering procedure removed redundant and/or irrelevant structural information (*e.g.* composition shift, sulfate losses) of the assignments. As a result, each selected assignment represented an independent observation of a sub-structure of the precursor. Note that the number of acetate groups, instead of sulfate groups, was included for clustering assignments. This was due to the ambiguity of modification number caused by acetate group. For example, in sequence #2 [1, 2, 3, 1, 6] (Figure 7), a peak assigned to $^{0,2}A_5$ (1Ac + $n SO_3$) can be equivalently assigned to B_5 (0Ac + $n SO_3$),

while a peak assigned to C_5 (1Ac + $n$$SO_3$) can be assigned to $^{2,4}A_6$ (0Ac + $n$$SO_3$). These ambiguous assignments had the same candidate Ac sites but different Ac numbers, and could not be differentiated from the mass value. In contrast, assignment ambiguity rarely involved the number of sulfate groups, but loss of sulfate group during dissociation may lead to misunderstanding of the sequence. By selecting the assignment with the highest number of sulfate groups within a cluster, the risk of misinterpretation caused by sulfate loss was reduced.

Another clustering procedure in HS-SEQ to reduce the risk of sulfate-loss will be described in step 4.

For each modification type, the algorithm went over step 2 – step 4 to construct the corresponding modification distribution.

2. *Estimating data ambiguity.* Only terminal assignments were considered for building the sequence, but internal assignments were used for estimating the risk of mis-assignment. For each modification type, the likelihood of correctly assigning a peak as an terminal assignment is assessed via uniqueness value $p(\mathrm{A})$, defined as

$$p(\mathrm{A}_k) = \frac{t_k}{\sum_{i \in \mathbf{S}} t_i}, \quad k \in \mathbf{S}, \text{ and } 0 \le t_i \le 1 \qquad \text{(Eq. 1)}$$

whereas **S** is the indices of isomeric assignments (assignments that correspond to the same mass value, k is the index of a terminal assignment in **S**, t_k is the weight associated with the cleavage type (*e.g.* glycosidic-bond cleavage, cross-ring cleavage, and the combinations thereof) of assignment k. For assignments of either glycosidic-bond

cleavage or cross-ring cleavage, t_i was set to 1.0; for assignments of internal cleavage type (except double cross-ring cleavage, *e.g.* AX ion), t_i was set to 0.2; for double cross-ring cleavage, t_i was set to 0 by default, which means assignments of double cross-ring cleavage were ignored. Note that the denominator of the uniqueness value (Eq. 1) was calculated based on assignments with non-redundant information of modification sites and modification number. If assignments of a mass value caused no additional ambiguity (Figure 5B) regarding the modification sites and modification numbers, they would not be counted into the denominator of the uniqueness value (Eq. 1).

3. *Constructing the assignment graph.* Terminal assignments were organized via an assignment graph model for each modification type (Figure 6B). Let X denote the set of candidate modification sites of an assignment and S denote the number of sulfate/acetate groups of the same assignment. Node *i* is a *parent* of node *k*, if $X_i \supset X_k$ and there is no such node *m* that $X_i \supset X_m \supset X_k$. Conversely, node *k* is a child of node *i*. Note that it is possible for a node to have more than one parent/child. For example, nodes representing $^{2,4}A_3$ and $^{3,5}A_3$ are both parents of the node representing B_2 with respect to the sulfation sites. Node *i* is defined as the *complement* node of node *k* if $X_k \cup X_i = \Omega$, $X_k \cap X_i = \varnothing$ and $S_k + S_i = S_{\text{precursor}}$, whereas Ω denotes the modification sites of the precursor sequence, $\varnothing$ denotes the empty set and $S_{\text{precursor}}$ is the total modification number of the precursor sequence. Conversely, node *k* is also a complement node of node *i*, and the two nodes are *complementary* to each other.

As shown in Figure 6B, the computation began with two connected dummy nodes, one representing the candidate modification sites and modification number of the precursor ion, and the other representing a virtual node with null modification sites and modification number of 0. For any new assignment, there was always at least one parent and one child in the graph. The terminal assignments were sequentially inserted into the graph by looking for their respective parent and child in the graph, and the insertion order depended on their respective confidence values. The confidence of an assignment relied on several factors: the assignment ambiguity, represented by the uniqueness value $p(\mathrm{A})$ (Eq. 1), and the assignment compatibility with the parent, child and complement node (if applicable).

The compatibility of assignment *k* was given by:

$$p(\mathrm{I}_k) = \begin{cases} C^{\min(D_k^G, D_k^L, D_k^U)}, & \min(D_k^G, D_k^L, D_k^U) \geq 0 \\ 0. & \min(D_k^G, D_k^L, D_k^U) < 0. \end{cases} \quad \text{(Eq.2)}$$

where *k* is the index of a terminal assignment, and *C* is a constant within [0,1] (0.8 by default). The boundary of the number of sulfate/acetate groups an assignment can carry is mathematically constrained by its parent, child as well as the complement assignment. D^G, D^L, and D^U represent the distances of the modification number of an assignment to the maximal status from different aspects. Constant *C* regulates the impact of the distance on the assignment compatibility. The rationale behind this formula is that the closer the modification number of an assignment is to the maximum, the more likely the node is to retain the original modification number information. This is especially useful for determining the sulfate distribution since we hope to select assignments that are most

likely to retain the intact sulfation information. D^G , D^L , and D^U in (Eq.2) are given by

$$D_k^G = S_{\text{total}} - (S_k + S_{k'}) \tag{Eq.3}$$

$$D_k^L = \begin{cases} S_{k_c} + (N_k - N_{k_c}) - S_k, & S_k \geq S_{k_c} \\ -1 & S_k < S_{k_c} \end{cases} \tag{Eq.4}$$

$$D_k^U = \begin{cases} S_{k_p} - S_k, & S_{k_p} - S_k <= N_{k_p} - N_k \\ -1, & S_{k_p} - S_k > N_{k_p} - N_k \end{cases} \tag{Eq.5}$$

where *k* is the index of the to-be-inserted node, k_c is the index of the child of node *k*, k_p is the index of the parent of node *k*, *k'* is the index of the complement node of node *k*, *S* is the modification number, and *N* is the number of candidate modification sites. If no complement node of node *k* has been recorded in the graph, let $S_{k'} = 0$. D^G is the distance inferred from a pair of complementary nodes (*i.e.* distance for **g**olden pair), D^L represents the distance inferred from the child node (*i.e.* distance for the **l**ower bound), and D^U is the distance inferred from the parent node (*i.e.* distance for the **u**pper bound).

In the event that more than one candidate parent / child node was present, the most confident one was selected, and its distance values were calculated accordingly.

In order to expand the assignment graph, terminal assignments were sorted in a descending order by their uniqueness values $p(\mathrm{A})$ (Eq. 1). In each cycle, assignments with the highest uniqueness values were retrieved. Note that for sulfation distribution, assignments with the same uniqueness value and same modification sites underwent an additional clustering procedure (as described in step 1) so that only the assignment with

the highest number of sulfate groups in each cluster was selected. This clustering procedure is useful for grouping nearby assignments covering the same candidate sulfation sites, *e.g.* B_3 and $^{1,5}A_3$ (see Figure 5B).

The compatibility (Eq.2) of assignment *k* involved two cases: 1) the compatibility with assignments in the current assignment graph (*background assignments*), denoted as $p(I_k^{bg})$, and 2) the compatibility with assignments of the same uniqueness values (*peer assignments*), denoted as $p(I_k^{peer})$. A virtual complement node *k'* of node *k* was forged (if not available), and the confidence value λ of node *k* was then calculated by:

$$\lambda_k = p(A_k) \times p(I_k^{bg}) \times \max(p(I_k^{peer}), p(I_{k'}^{peer})) \quad \text{(Eq.6)}$$

whereas *k* is the index of the current node, *k'* is the index of the complement node of node *k*, and $p(I_k^{peer})$, $p(I_{k'}^{peer})$ represents the compatibility value of node *k*, node *k'* in the context of peer assignments (Eq.2), respectively. Assignments with confidence value of 0 were ignored. The remaining peer assignments were sorted by their confidence values λ (Eq.6) in a descending order, and sequentially inserted into the assignment graph. Each insertion of a node was accompanied with the insertion of its complement node (either virtual or not).

4. *Updating modification distribution.* With only two dummy nodes in the graph, the modification was equally likely to occur on all possible modification sites along the precursor sequence. The insertion of a node into the assignment graph led to an update of a local region of the modification distribution (Figure 6B). From the perspective of the assignment graph, the confidence value λ (Eq.6) was responsible for arranging the order

of nodes for insertion into the graph; from the perspective of modification distribution, λ directed the way of updating the distribution. Intuitively, if the confidence value was close to 0, the inserted node had almost no impact on the current modification distribution; if it was close to 1, the modification distribution was updated based on the exact modification number of the assignment (Figure 4D). Therefore, λ controlled the effect of modification number of an assignment on updating the modification distribution (Figure 6B). The adjusted or "effective" modification number S' for assignment *k* was given by:

$$S'_k = \lambda_k S_k + (1-\lambda_k) S_k^{\mathrm{bg}} \quad \text{(Eq.7)}$$

whereas *k* is the index of the inserted node, *S* denotes the modification number, λ is the confidence value (Eq.6), and S_k^{bg} is the background modification number for node *k*, given by:

$$S_k^{\mathrm{bg}} = S'_{k_c} + L^{\mathrm{ori}}_{k_p - k_c} \times (N_k - N_{k_c}) \quad \text{(Eq.8)}$$

where *k* is the index of the inserted node, k_c is the index of the child of node *k*, S' is the effective modification number, *N* is the number of candidate modification sites and $L^{\mathrm{ori}}_{k_p - k_c}$ is the average modification number over the candidate modification sites sandwiched by the parent and child of node *k*. If the child node or parent node is the dummy node, let $S' = S$.

$L^{\mathrm{ori}}_{k_p - k_c}$ was given by:

$$L^{\mathrm{ori}}_{k_p - k_c} = \frac{S'_{k_p} - S'_{k_c}}{N_{k_p} - N_{k_c}} \quad \text{(Eq.9)}$$

whereas S' denotes the effective modification number, N is the number of modification sites, k_p is the index of the parent of node k, k_c is the index of the child of node k.

After the insertion, the original modification distribution was updated in the following way: for the subregion demarcated by the node k and its child k_c, the updated average modification number on each candidate modification site, or local modification likelihood, was given by:

$$L^{\text{ori}}_{k-k_c} = \frac{S'_k - S'_{k_c}}{N_k - N_{k_c}} \qquad \text{(Eq.10)}$$

Similarly, for subregion demarcated by node k and its parent node k_p, the updated local modification likelihood on each candidate site was also given by:

$$L^{\text{ori}}_{k_p-k} = \frac{S'_{k_p} - S'_k}{N_{k_p} - N_k} \qquad \text{(Eq.11)}$$

As more assignments were inserted into the assignment graph, the modification distribution was sliced into smaller pieces of subregions with the candidate modification sites and adjusted modification numbers specified.

5. *Organizing sequencing tasks.* The acetylation positions were identified first by selecting the most likely candidate acetylation sites based on predicted acetylation distribution. All assignments that reported inconsistent acetate numbers were removed in order to improve the accuracy of predicting sulfation distribution.

Ideally, the presence of cross-ring cleavage product ions facilitates locating the sulfate groups within each residue. In practice, however, cross-ring cleavage product ions were more likely to be associated with sulfate loss. As a result, HS-SEQ might

incorrectly distribute the sulfate group(s) within the residue. Fortunately, once the number of sulfate groups on each residue was determined, it was usually possible to deduce the priority of sulfation positions within residues based on the biosynthetic rules of HS.

3.5 Evaluation with Pure Standards

We tested HS-SEQ with 25 tandem mass spectra (Table I) from 9 synthetic HS sequences (Figure 7) on a desktop PC Intel i5 CPU (3.20GHz) and 16 GB memory. Those sequences are all known standards and widely used by multiple labs for HS study.

Sequence (charge)	*m/z*	# monoisotopic peaks	# assigned masses	Time(s)
#1 (4-)	375.72961	238	127	2.042
#1 (5-)	300.3823	238	98	1.951
#1 (6-)	250.15072	171	99	1.938
#2 (3-)	510.01074	175	93	6.905
#2 (4-)	382.2564	164	92	6.898
#2 (5-)	305.60354	284	186	6.994
#2 (6-)	254.50175	155	110	6.943
#3 (3-)	536.66297	97	62	7.224
#3 (4-)	402.24564	98	60	7.221
#3 (5-)	321.59491	194	129	7.547
#3 (6-)	267.8279	80	54	7.263
#4 (5-)	600.50781	277	108	73.111
#4 (6-)	500.4231	402	193	73.473
#4 (7-)	428.79002	407	194	74.576
#4 (8-)	375.06593	398	205	73.729
#5 (3-)	521.06975	129	59	3.105
#5 (4-)	390.55076	102	50	3.166
#6 (6-)	393.48524	156	74	8.362
#7 (3-)	391.3546	196	80	0.936
#7 (4-)	293.26412	190	107	0.951
#8 (2-)	547.5571	54	11	0.826
#8 (3-)	364.7023	81	36	0.826
#8 (4-)	273.27492	97	50	0.843
#9 (3-)	391.35453	140	67	0.921

#9 (4-)	293.26412	187	93	0.936

Table I Spectra information of the 25 spectra. Note that the calculation time includes only the time of generating theoretical fragments and modification distributions.

3.6 Comparison with Naïve Methods

We further compared the average ranks of the true sequences using all three methods (Table II), and the plotted the results in Figure 8A. A low rank value indicated high rank performance. The results suggested that the coverage method was the most stable among the three, and performed best for sequence #2 (charge 3-, 4-, 5- and 6-), #6 (charge 6-) and #7 (charge 3-). The performance of HS-SEQ (Cost) was close to the coverage method in general, with the exception that it lagged behind for sequence #1 (charge 4-, 5- and 6-), #2 (charge 5- and 6-) and #6 (charge 6-), while it targeted sequence #4 (charge 6-) accurately. The GP method performed poorly for #8 (charge 2- and 3-), #2 (charge 3-, 4-, 5- and 6-), #6 (charge 6-) and #4 (5-, 6-, 7- and 8-).

Sequence	Coverage	GP	HS-SEQ (Cost)	M_Coverage	M_GP	HS-SEQ (M_Cost)
#1 (4-)	3.5	2	12	1	1	1
#1 (5-)	3.5	2	75	1	1	4
#1 (6-)	3.5	3.5	16	1	1.5	1
#2 (3-)	18.5	130.5	26.5	2.5	14.5	2
#2 (4-)	9.5	45.5	13.5	1.5	8.5	1
#2 (5-)	9.5	36.5	29.5	1.5	6.5	1
#2 (6-)	9.5	45.5	114.5	1.5	4.5	1
#3 (3-)	3.5	10.5	17.5	1.5	2.5	2
#3 (4-)	9.5	12.5	3.5	2.5	3.5	1
#3 (5-)	3.5	6.5	7.5	1	2.5	1
#3 (6-)	12.5	22.5	12.5	2.5	3.5	1
#4 (5-)	18.5	32.5	15.5	1.5	4.5	1

#4 (6-)	6.5	42.5	1	2	5.5	1
#4 (7-)	86	288.5	70.5	3	15.5	2
#4 (8-)	602	8299.5	1630.5	15.5	727.5	3
#5 (3-)	5	17	8	1	5	1
#5 (4-)	2.5	2.5	2	1.5	1.5	1
#6 (6-)	9.5	60.5	30.5	1	4.5	1
#7 (3-)	8	12.5	15.5	2	2.5	1
#7 (4-)	1.5	6.5	1.5	1	2.5	1
#8 (2-)	5	35.5	6	1	7.5	1
#8 (3-)	2	8.5	2.5	1	2.5	1
#8 (4-)	1.5	2	1.5	1	1	1
#9 (3-)	6.5	11	13.5	2	4.5	3
#9 (4-)	1.5	5	2	1	1.5	1

Table II The average ranks of the true HS sequences using all methods. If two sequences have equal scores, fractional ranks – the mean values of their ordinal ranks ("1234" ranking) are used ("1 2.5 2.5 4" ranking).

The Z-score test (Figure 8B) provided a measure of the distinction of a true sequence from the background candidates. HS-SEQ (Cost) performed best for sequence #5 (charge 4-) and #4 (5-, 6-, 7- and 8-). The GP method performed best for #1 (4-, 5- and 6-), but produced no results for #8 (charge 2-) due to missing observations of golden pairs. The coverage method, the best performer in the average rank test, gave a mediocre performance in the Z-score test. This was not surprising, since ambiguous assignments in HS tandem MS spectra were expected to boost the scores of some candidate sequences. The results showed that HS-SEQ (Cost) gave good Z-scores consistently, especially for sequence #7 (4-, 56 candidates), sequence #9 (4-, 56 candidates), sequence #8 (3-, 70 candidates), sequence #5 (3- ~ 4-, 286 candidates), #3 (4-, 990 candidates), #6 (6-, 4,368 candidates) and #4 (5- ~ 8-, 1,381,380 candidates).

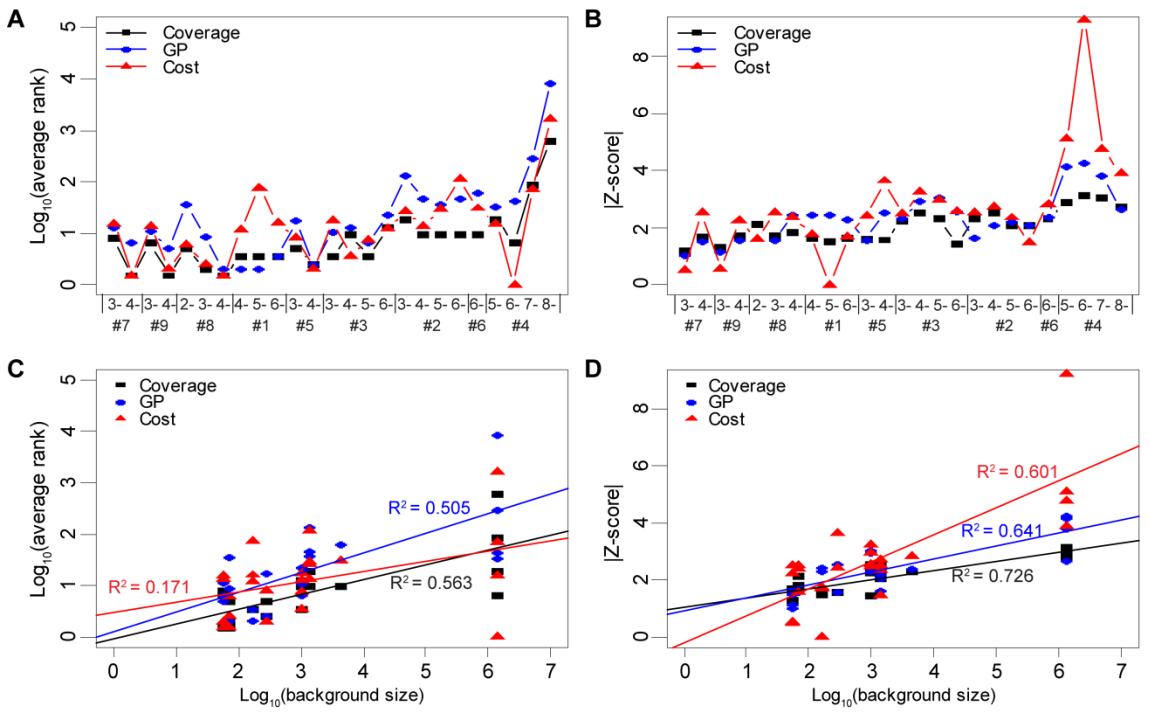


Figure 8 Comparison of HS sequencing methods. The performance for coverage method (denoted in black), GP method (denote in blue), HS-SEQ (Cost) (denoted in red) were compared using the 25 NETD spectra. (A) Comparison of the average ranks. (B) Comparison of the absolute values of Z-scores. (C) Comparison of correlations between average rank and background size. (D) Comparison of correlations between |Z-score| and background size. Note that in (A) and (B), the sequences were sorted in an ascending order by their background size.

We also examined the linear correlation between average rank and background size (Figure 8C). The results showed that the linear correlation was weak for HS-SEQ (Cost) (R^2 = 0.171), compared to strong correlation for the coverage (R^2 = 0.563) and GP (R^2 = 0.505) methods. It was interesting for HS-SEQ (Cost) to have significantly smaller

R^2 value, since a larger background size typically indicates a higher chance for a sequencing algorithm to make mistakes. For sequences of the same background size, HS-SEQ (Cost) tended to generate average ranks with higher variance compared to the naïve methods. This suggests that factors other than the background size may strongly affect the average rank performance of HS-SEQ (Cost). The regression line for HS-SEQ (Cost) (Figure 8C) had a smaller slope compared to other methods, which also suggests that the average rank from HS-SEQ (Cost) was less affected by the background size. In contrast, all three methods showed strong linear correlations on the relationship between the Z-score and the background size (R^2 = 0.726 for the coverage method, R^2 = 0.641 for the GP method and R^2 = 0.601 for HS-SEQ (Cost)), but the regression line for HS-SEQ (Cost) again was steeper than the lines for the naïve methods. This indicates that HS-SEQ (Cost) had special advantage in identifying the true sequence from the background, especially in the case of large background size. The characteristic distinctiveness of HS-SEQ was highly favorable for automatic HS sequencing.

Close examination of the modification distributions from HS-SEQ (Cost) provided an explanation. Some degrees of sulfate losses were observed with all dissociation methods, including NETD (116). This was more likely to happen for cross-ring cleavages. While the coverage method considered product ions with sulfate loss, the GP and HS-SEQ (Cost) methods required the presence of product ions with no sulfate loss. Fortunately, the total number of sulfate groups on each residue was usually well maintained in the sulfation distribution generated by HS-SEQ, because glycosidic-bond cleavages surrounding each residue were more likely to retain all sulfate groups. In this

sense, by ignoring the relative sulfation positions within GlcN residues, we expected HS-SEQ to perform well even in the average rank test.

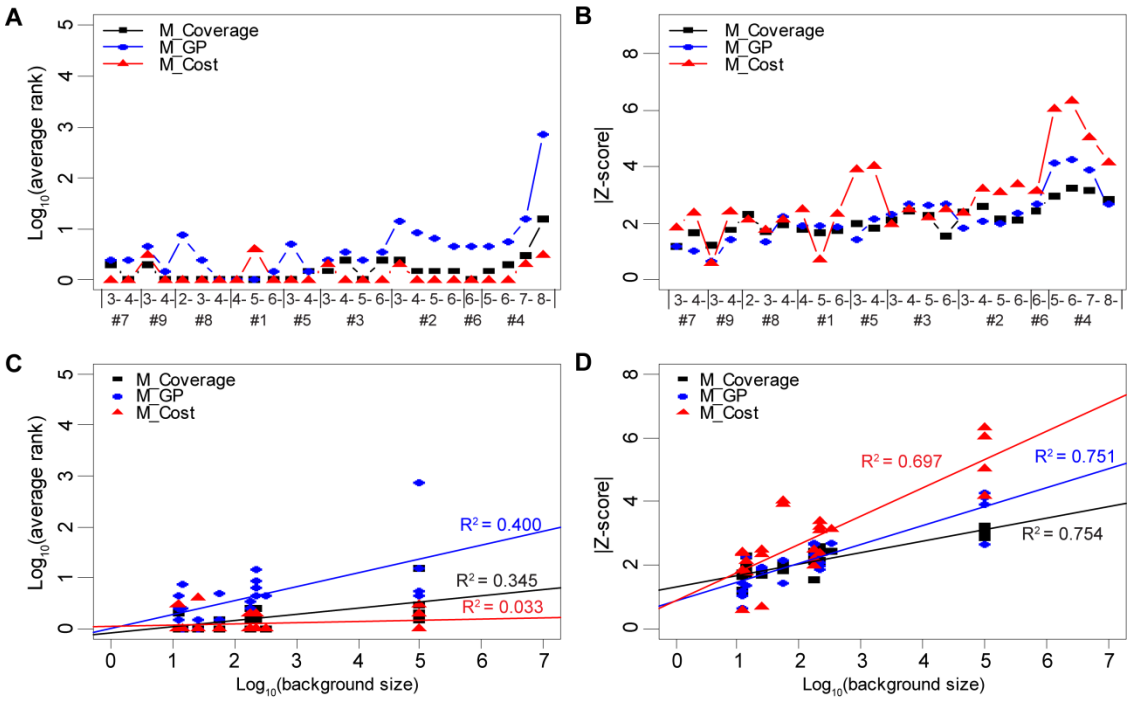


Figure 9 Comparison of updated version of HS sequencing methods. The performance for updated version of the coverage method (M_Coverage, denoted in black), GP method (M_GP, denoted in blue) and HS-SEQ (M_Cost, denoted in red) were compared using the 25 NETD spectra. (A) Comparison of the average ranks. (B) Comparison of the absolute values of Z-scores. (C) Comparison of correlations between average rank and background size. (D) Comparison of correlations between |Z-score| and background size. Note that in (A) and (B), the sequences were sorted in an ascending order by their background size.

We compared the updated version of all three methods (Figure 9). As expected, the HS-SEQ (M_Cost) method showed great improvement in the average rank test

(Figure 9A) and maintained its excellent Z-score performance (Figure 9B). Although the M_Coverage method provided comparable performance in the average rank test for most sequences, its suboptimal performance in the Z-score test made it less practical. As shown in Table II, HS-SEQ (M_Cost) successfully identified the correct structure as rank 1 for 19 out of 25 (76%) spectra, while M_Coverage identified 11 (44%) and M_GP identified 3 (12%). Since nearly all precursor sequences were present in multiple charge states, the supports of each precursor sequence from multiple spectra (each spectrum corresponds to a charge state) were also examined. We defined that if at least 50% of the spectra of the same sequence supported the true sequence as rank 1, then the sequence was correctly identified. The results (Table II) showed that HS-SEQ (M_Cost) correctly identified all sequences (100%). In contrast, M_Coverage deduced the correct structure for 66.7% of the sequences (sequence #1, #5, #6, #7, #8 and #9), and M_GP only worked for sequence #1 (representing 11% of the sequences).

The updated methods showed similar results in the correlation study (Figure 9C) and (Figure 9D) as the original methods (Figure 8C and Figure 8D). In the average rank vs. background size test, the linear correlation dropped for all three methods (for HS-SEQ, R^2 changes from 0.171 to 0.033; for the coverage method, R^2 from 0.563 to 0.345; for the GP method, R^2 from 0.505 to 0.400), but the distinction of R^2 between HS-SEQ (M_Cost) and the other two remained. The linear correlations between the Z-score and the background size were similar for all three methods (R^2 = 0.697 for HS-SEQ (M_Cost), R^2 = 0.754 for M_Coverage and R^2 = 0.751 for M_GP), albeit the regression

line for HS-SEQ (M_Cost) had the steepest slope. This suggests that as the background size grew, HS-SEQ (M_Cost) tended to provide the most distinctive results.

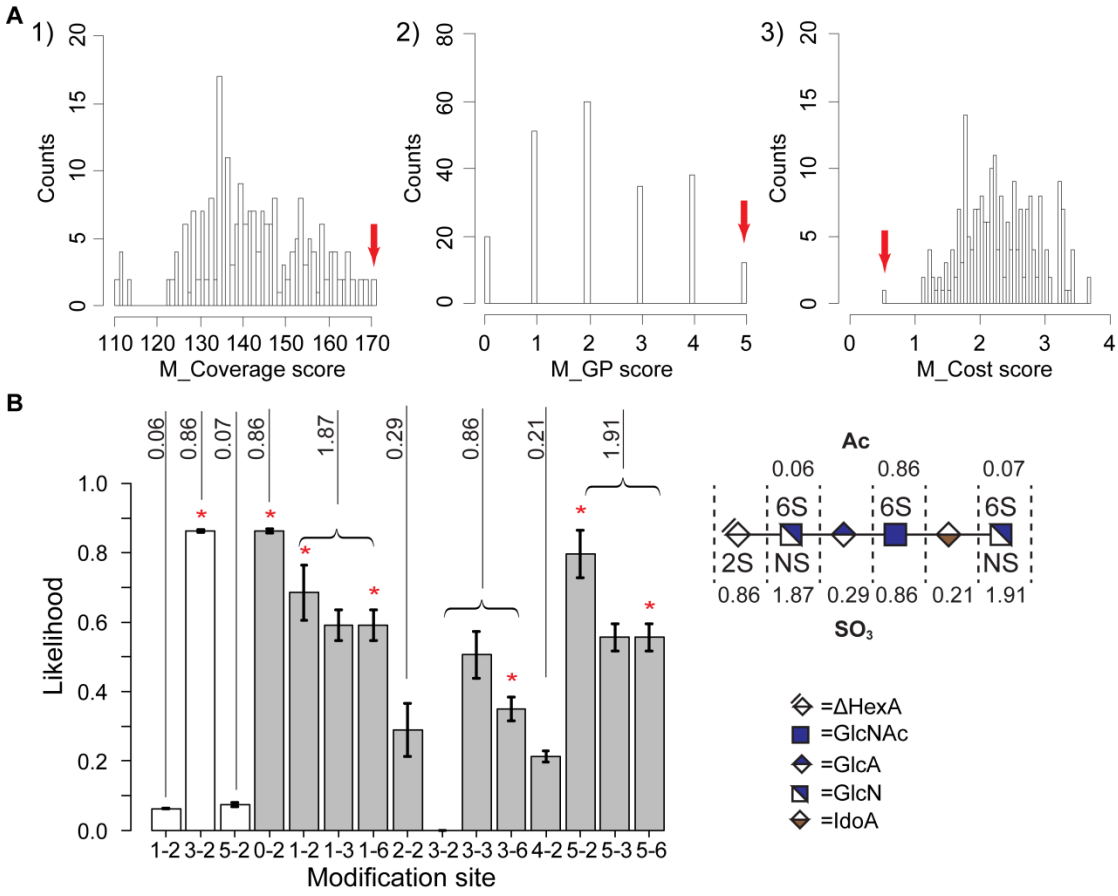


Figure 10 Example demonstrating the performance of HS-SEQ. (A) Comparison of histograms of candidate sequence scores using different methods. The calculation was based on tandem mass spectrum from sequence #2 (charge 5-). Red arrow flags the score of the true sequence structure. (B) Integration of results from multiple charge states. The modification distributions (bottom left) were calculated using data from sequence #2 (charge 3- ~ 6-). The modification number on each residue was then mapped to the original oligosaccharide sequence (bottom right). White bar denotes acetylation distribution, grey bar denotes sulfation distribution, and the error bar indicates standard

error. Digits beside the vertical solid lines represent the estimated modification number on each residue. Red asterisk indicates the positions where modifications actually occur.

When tied scores were assigned the minimum rank values, the original versions of the coverage and GP methods assigned the true sequences as rank 1 for most of the tested spectra. However, when tied scores were assigned the maximum rank values (data not shown here), all three methods performed poorly in identifying the true sequences. Histograms demonstrating the distinctiveness of three methods are given in Figure 10A.

To summarize, HS-SEQ (M_Cost) consistently provided the best performance regarding the average rank and distinctiveness, which enabled confident and high-throughput HS sequencing using NETD tandem mass spectrometry techniques.

3.7 Generation of Top Candidates

The generated modification distribution can be easily converted to a list of top candidate sequences. The underlying idea of the implementation is that for any candidate HS sequence *l* (except for the worst one), the best suboptimal sequences with respect to sequence *l* can be generated directly with no enumeration of all candidate sequences.

Take sulfation distribution for example. The distribution in HS-SEQ is expressed as an ordered list of digits showing the likelihood of sulfation across all potential sulfation sites. For precursor sequence with *n* sulfate groups, the top *m* candidate sequences can be generated in the following steps (Figure 11).

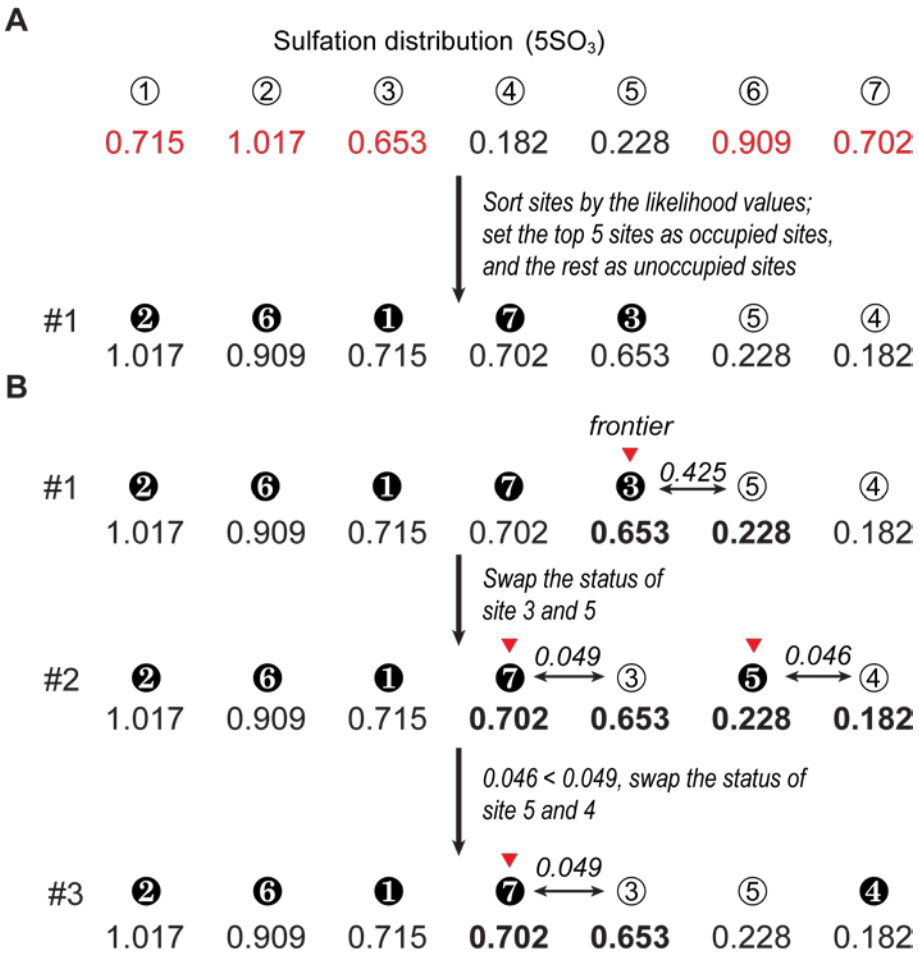


Figure 11 Generating top candidate sequences from modification distribution. (A) Generating the optimal candidate sequence from the modification distribution. For an HS sequence with *n* SO_3 groups, the optimal candidate #1 is selected by setting the top *n* candidate sites which are closest to 1 as "occupied" sites, while setting the rest as "unoccupied" sites. (B) Deducing the suboptimal candidate sequences. All the candidate sites in the optimal sequence #1 are sorted descendingly by their likelihood values. The *n*th occupied site is flagged as "frontier". The suboptimal sequence #2 is obtained by swapping the occupation status between the frontier site (site 3) and its unoccupied neighbor (site 5). The occupied sites which sit next to unoccupied sites are set as frontier

sites for sequence #2 (site 7 and 5). The next suboptimal sequence is obtained by choosing the frontier site which incurs the lowest likelihood decrease once swapping (0.049 for swapping the status of site 7 and 3, 0.046 for swapping the status of site 5 and 4) and taking the corresponding swapping action (swapping site 5 and 4). The black circle stands for the occupied site, and the white circle for the unoccupied site. The red triangle represents the frontier site. The red digit represents the likelihood values of the sites set as occupied for the optimal sequence #1, and the black bold digit represents the likelihood values of the sites involving in candidate swapping steps. The double ended arrow represents the potential swapping actions, and the digit above the arrow indicates the consequent likelihood decrease.

1. All sites are sorted by their sulfation likelihood values in a descending order. The statuses of the top *n* candidate sites are set as occupied, and the rest as unoccupied (Figure 11A). Since the number *n* is pre-defined in the precursor composition, this configuration (occupied/unoccupied) guarantees that the sum of the likelihood of all occupied sites (referred to as likelihood sum) is the largest among all choices. Note that the configuration of the complete candidate sulfation sites in fact represents the whole HS sequence. The configuration that produces the largest likelihood sum corresponds to the optimal sequence, which was denoted as *l*.

2. In sequence *l*, if an occupied site stays left to an unoccupied one, the former site is flagged as a *frontier*. Swapping the status between an occupied site and an unoccupied site maintains the balance of site numbers between occupied and unoccupied sites, and thereby generates an alternative sequence. Swapping the status between a

frontier and any unoccupied sites on its right decreases the likelihood sum, and the decreased value for this frontier is minimized when the swap occurred between the frontier and its right neighbor. There might be more than one frontier on sequence *l*, and each frontier is associated with a minimized decreases value through swapping with its right neighbor. Therefore, the status list for the best suboptimal sequence *l'* can be obtained by traversing all the swapping options between frontiers and their respective right neighbors, and selecting the option which causes the minimum decrease of the likelihood sum. Note that it is possible for sequence *l* to have more than one suboptimal sequence. The swapping process is illustrated in Figure 11B.

3. Take sequence *l'* as the new optimal sequence, and go to step 2 to find the next suboptimal sequence. Repeat the process until all *m* candidate sequences are generated.

Based on the NETD tandem spectra we tested, cross-ring cleavages had a large chance of losing sulfate groups. As a result, HS-SEQ might incorrectly identify the sulfation positions within GlcN residues but predict correctly for the total number of sulfate groups for the same residues. The selection of best swapping option discussed in step 2 might be adjusted according to the actual sulfate loss situation. For the NETD data that we worked on, swapping status between sites within the same residues may be preferred to swapping between sites that causes the minimum decrease of the likelihood sum but came from different residues.

3.8 Discussion

The framework of HS-SEQ can be envisioned as the model-view-controller (MVC) pattern in software design. In this sense, the modification distributions (view)

provided an intuitive representation of the results, the assignment graph (model) defined the relationship between peak assignments and mapped the relationship to modification distributions, and the confidence (controller) specified the priority of the peak assignments.

The HS-SEQ algorithm used a divide-and-conquer strategy to partition the sequence into smaller regions. No enumeration of candidate sequences was required in the process. The method was very efficient and required only a few seconds to generate the modification distributions from monoisotopic peak list, where the top candidate sequence list can be immediately deduced. Besides, tandem mass spectra from the same oligosaccharide sequence will generate modification distributions upon the same list of modification sites (for each modification type). This means it is very straightforward to integrate results from different charge states and dissociation methods (*e.g.* EDD and NETD) of the same precursor sequence.

Several assumptions were required for HS-SEQ to function well. The first was the acquisition of high-quality tandem mass spectra where most glycosidic-bond cleavages were present. Although ions with sulfate loss were usually present and tolerated by HS-SEQ, observation of product ions with no sulfate loss was necessary for successful structure identification. The second assumption required that a significant number of terminal containing product ions be unambiguously assigned. The best results were achieved for HS saccharides derivatized at the reducing end to break the structural symmetry. Data with low resolution may probably break the second assumption and lead to poor sequencing performance, in which case database searching can be a remedy.

Even with high-quality and low-ambiguity data, some problems remain under the current framework of HS-SEQ. One is that it did not contain a mechanism to resolve the conflicts of sulfation information from two assignments. For example, the sum of sulfate numbers from two complementary assignments may exceed the total number specified by the precursor, or a child assignment may contain more sulfate groups than its parent. The conflicts may arise from multiple events, such as sulfate loss, internal fragment disruption, random ion match, and co-existence of mixture. HS-SEQ simply removed the assignment with lower confidence value in order to resolve the conflict. This might not be proper for real samples. The heterogeneity and low abundance (missing information) of the species in real samples may break the assumptions of HS-SEQ and lead to misidentification of the HS sequences. Successful identification from real samples may require concurrent efforts from both the experimental part (extraction and separation) and the computational part (combination of *de novo* sequencing and database searching).

In conclusion, this chapter introduces HS-SEQ, a computational framework for high-throughput, accurate HS *de novo* sequencing. We expect that the method will apply to other GAG classes (21) since they all share similar chemical and structural properties. The divide-and-conquer strategy used in our method may also be instructive to the design of new high-resolution tandem MS sequencing algorithms for other complex molecules, once the sequencing problem and sub-problem have been well framed.

Chapter 4 *De Novo* Sequencing of Heparan Sulfate Mixture

4.1 Introduction

One of the most challenging problems in glycan identification is the heterogeneous nature of glycan molecules. This is especially the case for HS identification, due to the microheterogeneity of HS/heparin caused by nonrandom distribution of sulfate groups and IdoA residues.

We have shown that HS-SEQ can effectively discriminate the correct HS structure from up to millions of candidates (Figure 10A). Different from traditional tandem MS-based sequencing strategies, HS-SEQ shifts the focus from designing a powerful scoring function to reducing the ambiguity of individual peak assignments. If every critical peak can be unequivocally identified and annotated, the correct sequence can be deduced directly with no trouble.

In the context of noise interruption or even "chimeric spectra" (spectra from co-fragmentation of isomeric/isobaric precursor ions), accurate assignment of individual peaks becomes tremendously difficult. For pure standards, as discussed in Chapter 3, a single peak in the tandem mass spectra can be assigned to multiple assignments, and one assignment does not guarantee the rejection of alternative ones. If the sample contains more than one sequence, a single peak may map to multiple candidate sequences, which further increases the level of ambiguity. A database search-based solution is easy to implement, by looking for sequences with the highest matching scores. The simplicity in implementation comes at a price in performance: the possibilities of isomer candidates increase combinatorially as the chains become longer and the sulfation degree reaches to

a moderate level. In addition, as shown by the coverage method (Figure 10A) in **3.6 Generation of Top Candidates**, it is also likely that a database search method will end up with a large group of candidates in the top tier, which makes this type of methods less practical.

HS-SEQ (106) provides several guidelines that may benefit high-throughput HS sequencing for complex samples: 1) no exhaustive enumeration of theoretical isomer sequences; 2) in simple situations (two isomers with long common subsequences), the true isomers should be distinct from the background; and 3) in complex situations (multiple isomers, or isomers with almost no common subsequences), better null results than wrong prediction. Point 1 means rapid sequencing regardless of the length and complexity of the isomer sequences. Point 2 implies that the true isomers should be picked accurately and automatically with few false positives. Point 3 suggests that the results should also be indicative of data quality (quality control). Even the complexity of spectra prohibits the program from generating discriminant predictions, the results should still flag the situation and recommend extra efforts in sample separation.

Following the guidelines, we are working on the development of MULTI-HS-SEQ, which aims to assist the automatic identification of HS isomers from mixture samples. As mentioned before, each peak assignment specifies a set of candidate sulfation sites and the number of sulfate groups owned by these sites. Taking the constraint of the precursor into account, one peak assignment automatically suggests an assignment on the complementary region of the sequence. In this sense, we use the term *cleavage* to specify that each peak assignment corresponds to a complementary

assignment (it can be virtual). A simplified example of three isomers is given in Figure 12A assuming that one residue contains at most one sulfate group and no internal fragments. Note that sulfate loss is taken into account, so the only way to detect the existence of isomers is through the sum of sulfur numbers with two complementary peak assignments (Figure 12B). If their sum is beyond the constraint of the precursor composition, then it indicates that fragments from two isomers do not agree with each other. The cleavages can be organized in a graphic model (Figure 12C) where each node represents a cleavage and the edge specifies the relative positions (from NRE terminal to RE terminal) of the connected cleavages. An edge that uniquely points to or points from a cleavage is essential and has to be included in at least one path (Figure 12C). In the simplified example, with the law of parsimony, only two paths (red arrows) and two candidate sequences are sufficient to explain all cleavages. If the inferred cleavages are taken into account, at least three paths (consisting of red/blue arrows) are required and the candidate space can increase up to 6 sequences. The true isomers are between the minimum two sequences and the maximum 6 sequences.

However, when ambiguous assignments are taken into account, the graph corresponds to 16 different paths (isomers) although only two isomers are true (see Figure 13). Besides, if fragment undergoes complete sulfate loss, or complementary peak assignments are missing, the true isomer may even be included in the candidate sequences.

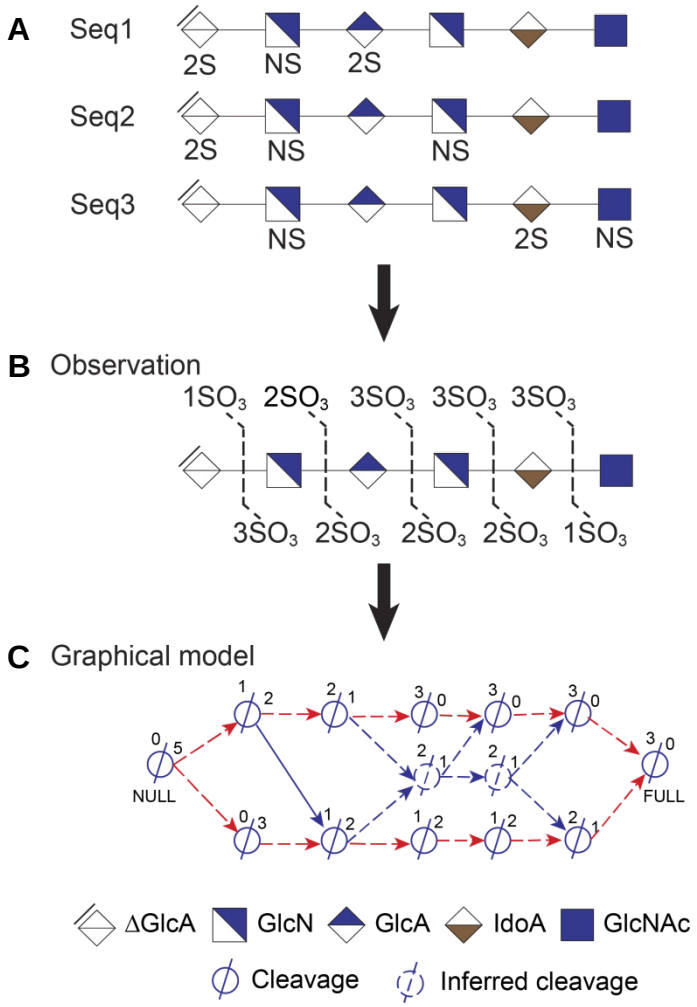


Figure 12 Simple model showing the graphical representation of cleavages and sequences. (A) Three isomers each with 3 sulfate groups. Each cleavage (red or blue line) divides the sequence into two parts. The numbers of occupied sites (sulfate groups) on each part is fixed. Black digit represents the observed number of occupied sites, and red digit represents the inferred number considering about sulfate loss. (B) Observation of sulfate groups indicated by each cleavage. For each cleavage, only the variant with the highest number of sulfate groups is recorded. (C) Sequence inference based on the graphical model. Each node represents a cleavage where the numbers specify the

numbers of sulfate groups on the non-reducing end and reducing end, respectively. Red dash arrow means the edge is essential (unique path). When inferred cleavage (dashed symbol) is considered, the corresponding essential edge is represented as blue dash arrow.

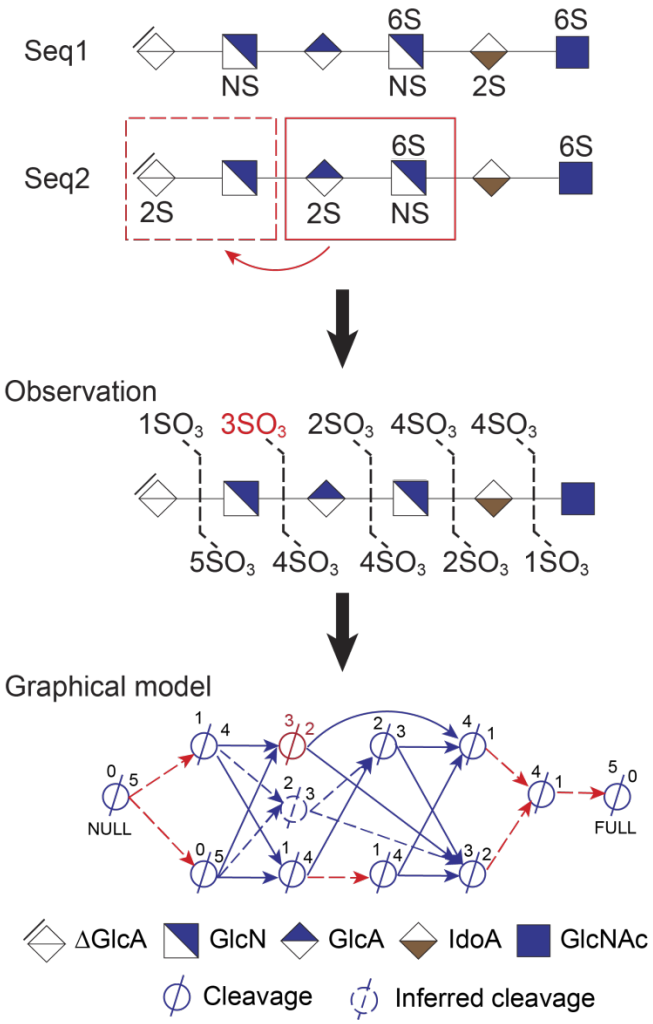


Figure 13 Graphic representation of cleavages from HS isomers in complex situation. When an internal fragment (solid square) carries more sulfate groups than its isomeric terminal fragment (dashed square), the sulfation information on this terminal fragment is overwritten, and the graph structure becomes complicated (16 possible paths

in the example). Red cleavage specifies the place where the sulfate information is mis-identified.

Therefore, in order to improve the identification performance for mixture sample, it is important to include more confident evidences regarding the distribution of sulfate groups. From the perspective of experimental design, chemical derivatization (117) converts the labile sulfate groups into stable acetate group (labeled) and removes the confidence issue caused by sulfate loss. The price is that protons and sodiums both can serve as charge carriers, which may potentially increase the assignment ambiguity. Another way is to look for high-confidence internal fragments that maintain the intact sulfation information.

4.2 Method Description

4.2.1 Data Preprocessing

Preprocessing of raw data and peak assignment followed the same procedure as described in Chapter 3. All the identified monoisotopic peaks were converted to the format of combination of backbone cleavage and modifications (106), *e.g.* $B_4 + 3SO_3 + 1Ac - 2H$. Future work will include a targeted method to iteratively (*e.g.* B_4, $B_4 + 1SO_3$, $B_4 + 2SO_3$, …) identify peak clusters with low abundance.

4.2.2 Sulfate Loss in Fragments

Internal cleavages seem to undergo significant sulfate loss, and are therefore not included by HS-SEQ for determining the sulfate positions. However, for HS

disaccharide analysis, internal cleavage of $^{0,2}A_2/Y_1$ type can be used for determining the occupation status of 2-*N* position on GlcN. It may also be the case for HS oligosaccharides. As shown in Figure 14, two different isotopic peaks (*m/z* 291.9812 *z* 3- and *m/z* 438.4755 *z* 2-) from Arixtra (6-) NETD tandem mass spectra suggest the existence of internal fragment $C_4/^{1,5}X_4$ + 4 SO_3. HS-SEQ program showed that there was no alternative assignment for this fragment under current fragmentation settings. There may be more intact internal fragments, but it is difficult to confirm their existence directly due to their low abundance and ambiguity in peak assignments. A confident confirmation requires supporting information from confident terminal assignments, *e.g.*, confining the unique highly sulfated local region so that all alternative assignments of the internal assignment are rejected. This is the strategy HS-SEQ uses to improve the confidence of terminal assignments, and should also be applicable to internal assignments.

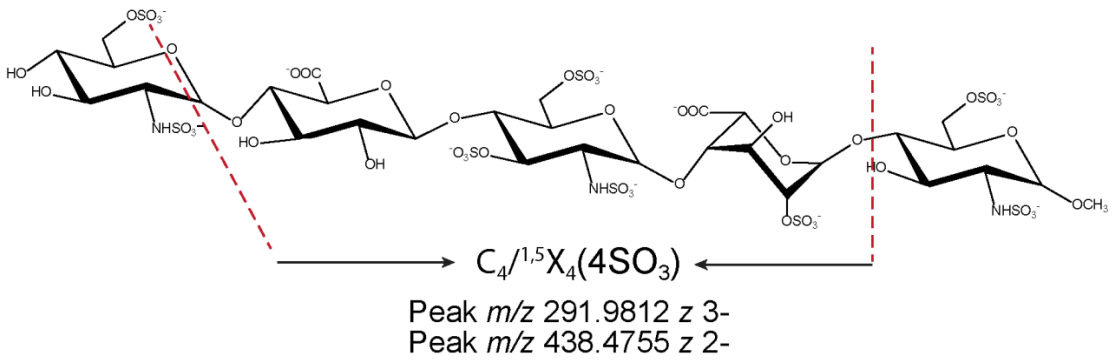


Figure 14 Example from Arixtra (6-) tandem mass spectra showing internal fragment without sulfate loss. The presence of the internal fragment was supported by two identified monoisotopic peak with different charge states.

Based on the limited data we have, it seems that the intact internal fragments tend to combine a cross-ring cleavage and a glycosidic-bond cleavage. One evidence is that in the NETD tandem mass spectra, peaks assigned to cross-ring cleavages are generally more abundant than glycosidic-bond cleavages and therefore are presumably easier to generate. Fragments from cross-ring cleavages may easily undergo further fragmentation to generate the internal fragments. However, the test results from HS-SEQ suggest that cross-ring cleavages are more likely to lose sulfate groups. In fact, that's the reason that we focused on the sum of sulfate groups for each residue instead of specific position inside the residue. On the other hand, internal fragments combining two glycosidic-bond cleavages (*e.g.* B/Y type and C/Z type) seems more likely to lose sulfate groups. This is good news, as internal fragments will have fewer chances to overwrite the sulfation information of a NRE-terminal fragment. Whenever we observe two complementary terminal assignments (not virtual) containing more sulfate groups than the constraint of the precursor, we tend to believe that it is a hint for isomer instead of internal fragmentation. Understanding of the fragmentation patterns can contribute to removing data ambiguity in tandem mass spectra, and eventually improve the sequencing results.

4.2.3 Principle of MULTI-HS-SEQ

Based on the discussion above, I propose an algorithm, named MULTI-HS-SEQ, to extend the current HS-SEQ framework to support the co-fragmentation of isomers. The model of MULTI-HS-SEQ is demonstrated using two HS oligosaccharide isomers of composition [1, 4, 5, 1, 8] (Figure 15A). Suppose both isomers produce an intact internal

fragment at the same cleavage sites. MULTI-HS-SEQ builds a graphical model (Figure 15B) based on filtered non-redundant peak assignments, as described in *3.4.3 Sequence Construction*. Note that each node in the graph represents a cleavage and a distinct way of distributing sulfate groups to the two cleaved ends, as indicated by the integers above the cartoon symbol. Each edge connects two nodes from the NRE terminal to RE terminal. One node can connect to multiple nodes if there are more than one way of distributing the sulfate groups.

If intact internal fragments are detected, they can contribute to minimizing the possible paths. In Figure 15B, an internal assignment containing 7 sulfate groups connects two nodes in the graph (red dashed arrow), and suggests that the nodes should be grouped into the same sequence. Based on the law of parsimony, another node pair (blue dashed arrow) should be grouped in a different sequence. This restriction effectively drops down the candidate paths from 16 to 8 (divided by 2). If more intact internal fragments can be detected and assigned to the right region, which may be hard, the final candidate paths can end up with a very small number.

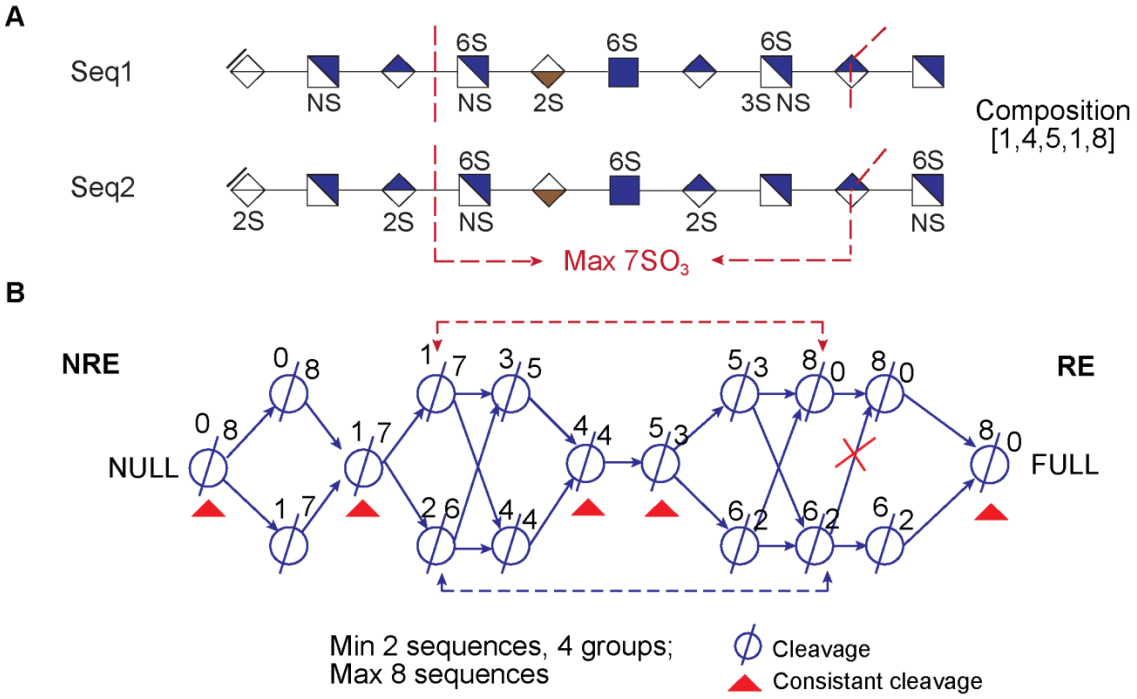


Figure 15 Model of MULTI-HS-SEQ in HS mixture sequencing. (A) Two HS isomers (virtual) with composition [1,4,5,1,8]. (B) Graphical model of MULTI-HS-SEQ. Note that internal fragment information (red and blue dashed arrow) can be used to reduce the candidate paths. Red triangle represents the consistent cleavages of the two isomers.

Each candidate path corresponds to a modification distribution, as discussed in Chapter 3. The mapping from modification distribution to candidate sequences depends on the quality and density of the cleavages. Missing cleavages (*e.g.* cross-ring cleavages) will lead to multiple candidate sequences on the top tier.

The topology of the graph also provides a way to evaluate the heterogeneity of the sample. If two isomers are aligned to each other, there will be several cleavage sites (red triangles in Figure 15B) where they share the same way of distributing sulfate groups.

The number of such cleavage sites (≥ 2) can be used to measure the heterogeneity of the isomers in the sample. If the number is just two (only NULL and FULL nodes), the isomers contain no common subsequences, which is the worst case for mixture sequencing. Fortunately, such isomers are more likely to be separated in LC step and no issue will be caused.

The MULTI-HS-SEQ program is currently under active development. The latest C++ source code can be downloaded from https://github.com/hh1985/multi_hs_seq.

4.3 Discussion

The inference of HS candidate sequences shares similarity with the protein inference problem in proteomics study. Specifically, a single peptide (peak assignment) can map to multiple proteins (HS isomers) while multiple peptides (peak assignments) may map to the same protein (HS isomer); observation of more peptides (peak assignments) adds confidence to the final identification. However, MULTI-HS-SEQ builds the relationship between peak assignments, which limits the expansion of candidate space. As a result, MULTI-HS-SEQ avoids the overhead on protein inference and makes no compromise between accuracy and speed.

It is important to reiterate the role of confident peak assignments in HS-SEQ/MULTI-HS-SEQ sequencing. HS-SEQ provides strategies to improve assignment uniqueness and look for assignments with intact sulfate groups. MULTI-HS-SEQ follows the basic idea. Further work includes implementing a targeted deconvolution-deisotoping algorithm to identify low-abundance monoisotopic peaks and collecting more

ions that have no sulfate loss. It is also important to design a function to infer the intact internal fragments due to their effect in minimizing candidate paths.

MULTI-HS-SEQ does not utilize the information of peak intensity. This simplifies the model, while also prevents the differentiation of IdoA from GlcA. Once the epimers undergo co-fragmentation, their fingerprints (distinct peak intensity pattern) are wiped out. From the perspective of experimental design, we expect that high-quality HPLC and new techniques such as ion-mobility can contribute to the separation of HS epimers. From the perspective of algorithm implementation, we hope the construction of HS library will serve to the exploration of fragmentation patterns for epimers.

In order to sequence HS isomers from natural samples, significant efforts are still required for sample preparation and generation of high-quality tandem mass spectra. In the near future, a typical workflow of exploring the structure-function relationship of HS could be: extracting and separating HS samples from interesting cells or tissues, generating NETD tandem mass spectra, predicting the candidate structures (using sequencing algorithm, *e.g.* MULTI-HS-SEQ), identifying epimer positions by searching against HS library (using database searching strategy), testing the response of specific HS sequence in *in vitro* experiments using chemoenzymatically synthesized HS oligosaccharides (110), and finally, validating the structure-function relationship with *in vivo* experiments. These endeavors, once achieved, will eventually contribute to deciphering the mysterious “glycan code” encoded in HS (16).

Chapter 5 Computational Pipeline for Heparan Sulfate Structure Identification

5.1 Architecture of Pipeline for Heparan Sulfate Sequencing

Interpretation of HS tandem mass spectra is a tedious and error-prone procedure, and largely requires user expertise and intensive manual inspection. This is especially the case when dense, noisy and low-abundant peaks are present in the tandem mass spectra. Even with successful data reduction and sequence identification, reporting and visualizing the information-rich spectra manually is still time-consuming.

Although data preprocessing such as noise reduction, peak picking, and deconvolution/deisotoping and results reporting seems like a routine task for any mass spectrometry-based pipeline, this is far from a solved problem for HS identification. Typically, fragments from HS precursors contain sulfate groups ranging from zero up to the maximum number of sulfate groups allowed on the fragments. The uncertainty of sulfate groups in a product ion doesn't only increase the difficulty in identifying the fine structure of HS, which has been handled by HS-SEQ, but also affects the quality of data preprocessing and post-processing. Therefore, programs automatically handling these tasks are highly demanding. In order to improve the productivity of HS sequencing, I designed several auxiliary modules with HS-SEQ as the core engine. The architecture of the complete pipeline is shown in Figure 16.

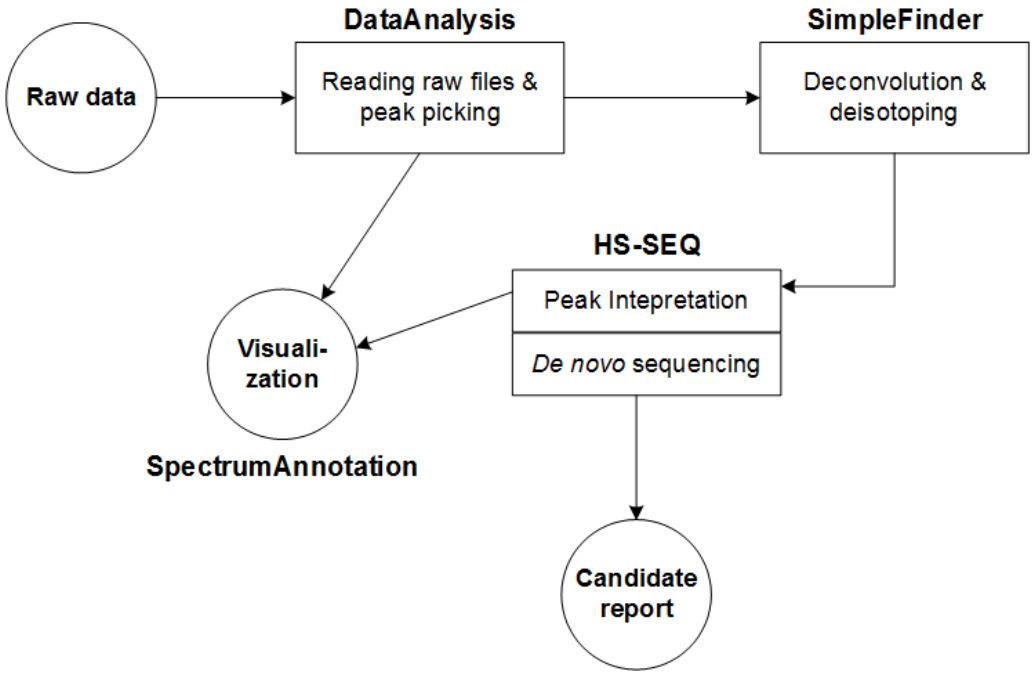


Figure 16 Workflow for automatic HS sequencing. Bruker software DataAnalysis is used to read the raw file and export peak list. A home-made deconvolution/deisotoping module SimpleFinder is used to convert the peak list into monoisotopic peak list with charge states identified. HS-SEQ finishes the peak interpretation and *de novo* sequencing, as described in Chapter 3. The results are either reported in a text file or visualized through visualization tool SpectrumAnnotation.

5.2 Deconvolution / Deisotoping

In mass spectral data analysis, deconvolution/deisotoping is an important preprocessing step that reduces data redundancy and selects peaks illustrating structural information. Deconvolution represents the procedure of separating overlapping isotopic clusters, while deisotoping is the process of removing redundant isotope peaks. Charge

state has to be assigned in the procedure in order to include all isotope peaks belonging to the same isotopic cluster. Low-abundance product ions may produce incomplete isotopic clusters. As a result, a program may either fail to assign the charge state (*e.g.* isotope peaks with low S/N ratio), or assign a wrong charge state (*e.g.* charge state 4- can be misidentified as 2- when A+1 peak is missing). Low accurate spectrum may cause the isotope peaks to be treated as monoisotopic peaks, which increases the chance of false-positive peak assignments.

In a sequencing procedure such as database search and naïve *de novo* sequencing, the spectrum interpretation is straightforward. One can list all theoretical fragments from the candidate sequence based on known fragmentation rules (101), and convert the elemental composition of each fragment into a theoretical isotopic cluster. A true isotopic cluster should match the theoretical one well given the error (*m/z* and intensity) range of the instrument and charge state inferred from the data. Deisotoping in bottom-up sequencing requires an approximated composition of the candidate isotopic cluster in order to generate theoretical isotopic distribution. The bottom line of the approach is that the estimated composition should produce an isotopic cluster similar to the experimental one. In proteomics, the most widely used approximation model is AVERAGINE (88), which uses an average amino acid with formula $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$ and molecular mass of 111.1254 Da. For HS sequencing, this is not the case. For two fragments with similar mass values, the isotopic cluster pattern of the high-sulfur fragment is dramatically different from the low-sulfur peer (Figure 17).

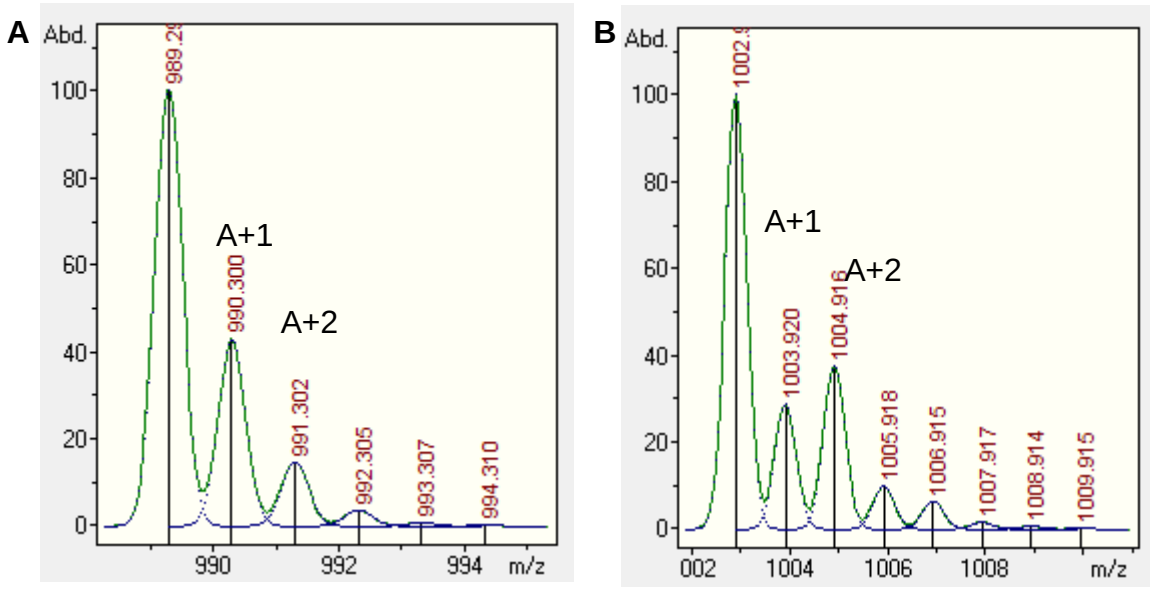


Figure 17 Comparison of isotopic clusters with close mass values but different sulfur contents. (A) Non-sulfur fragment composition $C_{37}H_{54}O_{28}N_3$. (B) High-sulfur fragment composition $C_{20}H_{30}O_{32}N_2S_6$. The relative abundance values of A+1 and A+2 peaks from the two compositions are distinct from each other.

In order to provide accurate monoisotopic peak list, I extended current AVERAGINE model so that the algorithm is able to tolerate fragments or molecules with varying sulfur content. I further implemented a C++ version of BRAIN (119), an algorithm that efficiently generates aggregated theoretical isotopic distribution from given elemental composition. For an isolated isotopic cluster, the processing consists of two steps: 1) identifying the charge state and generating the neutral mass value for the monoisotopic peak and 2) generating an approximated composition with optimized sulfur number. In cases there're suspicious overlapping isotopic clusters, the module contains

an extra step to iteratively extract the currently most abundant isotopic clusters in a greedy way (120).

5.2.1 Extended AVERAGINE Model for Generating Theoretical Composition for Heparan Sulfate Fragment

Compared with peptide fragments, HS fragments introduce an extra variable – the number of sulfate groups that alters the shape of the isotopic distribution. Therefore, it is important to expand a single AVERAGINE formula (118) to composite formulas (103) in order to improve the identification rate of isotopic distributions (121). The composition of an HS fragment can be decomposed into two parts: the non-sulfated HS backbone (including acetate groups) and the sulfate groups. The estimated composition is shown as follows:

$$\mathrm{HS}_{compo} = \text{neutral mass}/100 \times C_x H_y O_z N_m + \kappa \mathrm{SO}_3 \qquad \text{(Eq.12)}$$

whereas C_x =3.7238523, H_y = 5.4425534, O_z = 2.8645018, and N_m = 0.2864502, which represent the average atom numbers of carbon, hydrogen, oxygen (excluding oxygen atoms in sulfate groups) and nitrogen per 100 Da in a HS fragment. κ represents the number of sulfate groups carried by the fragment, which controls the impact of sulfate groups on the isotopic distribution. Note that in high resolution condition, the mass defect of sulfur element causes the splitting of A+n (n > 1) isotopic peaks and significantly affects the shape of isotopic distribution.

In low-resolution condition, for a given candidate isotopic cluster, all possible κ values are explored to find the optimal κ value, and the corresponding fragment composition is estimated by:

$$\mathrm{HS}_{compo} = (\text{fragment mass} - \kappa \times M_{SO_3}) / 100 \times C_x H_y O_z N_m + \kappa SO_3 \quad \text{(Eq.13)}$$

whereas M_{SO3} represents the mass value of a sulfate group.

In high resolution condition, the value of κ can be derived by dividing the abundance of the monoisotopic peak by the abundance of the A+2 peak.

$$\kappa = \mathrm{I}_A / (\mathrm{I}_{^{34}\mathrm{S}:A+2} \times \mathrm{A}_{^{34}\mathrm{S}}) \quad \text{(Eq.14)}$$

whereas I_A is the abundance of the monoisotopic peak (A peak), $\mathrm{I}_{^{34}\mathrm{S}:A+2}$ is the abundance of the A+2 ^{34}S-containing peak, and $\mathrm{A}_{^{34}\mathrm{S}}$ is the natural relative abundance of ^{34}S (~4%). The estimated fragment composition can then be given by (Eq.13)

5.2.2 C++ Implementation of the BRAIN Algorithm for Generating Theoretical Isotopic Distribution

In order to correctly interpret fragment peaks, one of the most critical steps is to identify their isotopic distributions from the tandem mass spectra. The shape of the isotopic distribution illustrates the position of the monoisotopic peak, which allows further structural identification. Calculation of the complete theoretical isotopic distribution, especially aggregated peaks each consisting of multiple isotope peaks that contribute the same amount of neutrons, is a combinatorial problem (122). Recently, Claesen *et al.* (121) proposed a generalized polynomial method to efficiently calculate

the exact center mass of aggregated peaks, which was included in the BRAIN (Baffling Recursive Algorithm for Isotopic distribution calculations) application (119) in R GNU.

The intensity distribution $Q(I;v,w,x,y,z)$ of a composition $C_vH_wN_xO_yS_z$ can be expressed as a polynomial:

$$\begin{aligned} Q(I;v,w,x,y,z) &= (P_{C_{12}}I^0 + P_{C_{13}}I^1)^v \\ &\times (P_{H_1}I^0 + P_{H_2}I^1)^w \times (P_{N_{14}}I^0 + P_{N_{15}}I^1)^x \\ &\times (P_{O_{16}}I^0 + P_{O_{17}}I^1 + P_{O_{18}}I^2)^y \\ &\times (P_{S_{32}}I^0 + P_{S_{33}}I^1 + P_{S_{34}}I^2 + P_{S_{36}}I^4)^z \\ &\equiv \{Q_C(I)\}^v \times \{Q_H(I)\}^w \times \{Q_N(I)\}^x \times \{Q_O(I)\}^y \times \{Q_S(I)\}^z \end{aligned} \quad \text{(Eq.15)}$$

whereas I represents the additional neutron number relative to the monoisotopic variant, v, w, x, y, z are the atom numbers of carbon (*C*), hydrogen (*H*), nitrogen (*N*), oxygen (*O*) and sulphur (*S*), $P_{C_{12}}, P_{C_{13}}, ..., P_{S_{36}}$ correspond to the natural relative abundance of the isotopes. (Eq.15) can be further converted into a polynomial function:

$$Q(I;v,w,x,y,z) = \sum_{j=0}^{n} q_j I^j \quad \text{(Eq.16)}$$

whereas q_j is the coefficient (relative abundance) of the A+*j* aggregated peak, and *n* is the sum of the atom numbers of all elements in the composition ($v+w+x+y+z$ in the example). It turns out that the coefficient q_j can be calculated iteratively by:

$$q_j = -\frac{1}{j}\sum_{l=1}^{j} q_{j-l}\psi_l \quad \text{(Eq.17)}$$

where parameter ψ_l can be calculated by:

$$\begin{aligned}\psi_l &= \sum (r^{-1})^l \\ &= v(r_C)^{-l} + w(r_H)^{-l} + x(r_N)^{-l} + y(r_O)^{-l} \\ &\quad + y(\overline{r}_O)^{-l} + z(r_{S,1})^{-l} + z(\overline{r}_{S,1})^{-l} + z(r_{S,2})^{-l} + z(\overline{r}_{S,2})^{-l}\end{aligned} \qquad \text{(Eq.18)}$$

whereas r_C, r_H, r_N is the unique roots of polynomial $Q_C(I)$, $Q_H(I)$, and $Q_N(I)$, while r_O, $\overline{r}_O$ and $r_{S,1}, \overline{r}_{S,1}, r_{S,2}, \overline{r}_{S,2}$ correspond to conjugate roots of polynomial $Q_O(I)$ and $Q_S(I)$.

Fernandez-de-Cossio Diaz and Fernandez-de-Cossio recently solved the polynomial specified by Claesen *et al.*(121) for the center mass using a fast Fourier Transform (FFT) approach (123), and implemented their new algorithm (referred as FTMC) in C#. They made a comparison between the BRAIN method and FTMC method in terms of their computational performance. BRAIN was originally implemented in R, which is an interpreted language and known for its low efficiency in computationally intensive work. To facilitate a direct comparison, I collaborated with Dittwald and Valkenborg, and ported the BRAIN method to C++ (94), since the performance of C++ and C# for standard bioinformatics algorithms were shown to be close (125).

We consider the following factors which may affect the numerical efficiency of the two programs: 1) algorithm complexity, 2) selection of heuristic methods, and 3) algorithm implementation & optimization.

Fernandez-de-Cossio Diaz and Fernandez-de-Cossio proposed an FFT-based method to solve the polynomial generating function, and showed that the time complexity of FTMC is of O(Nlog$N$) while the iterative formula in BRAIN has O(N^2), where N is the number of computed peaks. However, the actual performance of algorithms may also depend on factors other than time complexity and should be addressed carefully. For

example, the authors of FTMC method relates the number of computed peaks *N* to the mass *m* of the tested molecule. The heuristic suggested a close relationship between the performance of the algorithm and the molecular classes. For this reason, we proposed some heuristics that relate to the molecular mass and evaluated their impact on the algorithms' performance.

The original heuristic implemented by the BRAIN (R) application is shown as follows:

$$N = \max(\lceil 2 \times (\text{mass}_{average} - \text{mass}_{monoisotopic}) \rceil, 5) \quad \text{(Eq.19)}$$

As pointed out by Fernandez-de-Cossio Diaz and Fernandez-de-Cossio, this heuristic may not provide good coverage of the aggregated distribution for large molecules. In order to achieve a reasonable coverage of the aggregated distribution for large molecules, we proposed a heuristic to calculate the number of peaks:

$$N = \max(\lceil 2 \times (\text{mass}_{average} - \text{mass}_{monoisotopic}) \rceil, 50) \quad \text{(Eq.20)}$$

Note that this heuristic is chosen since it generates a reasonable number close to the average mass of a molecule. It was designed for comparison purpose and may have little value in practice.

Figure 18 compares the performance of heuristics used by BRAIN and FTMC. The result shows that FTMC requires the calculation of fewer peaks to achieve a high coverage of the cumulative abundance. It also suggests that both BRAIN and FTMC overestimates the number of required peaks for small molecules.

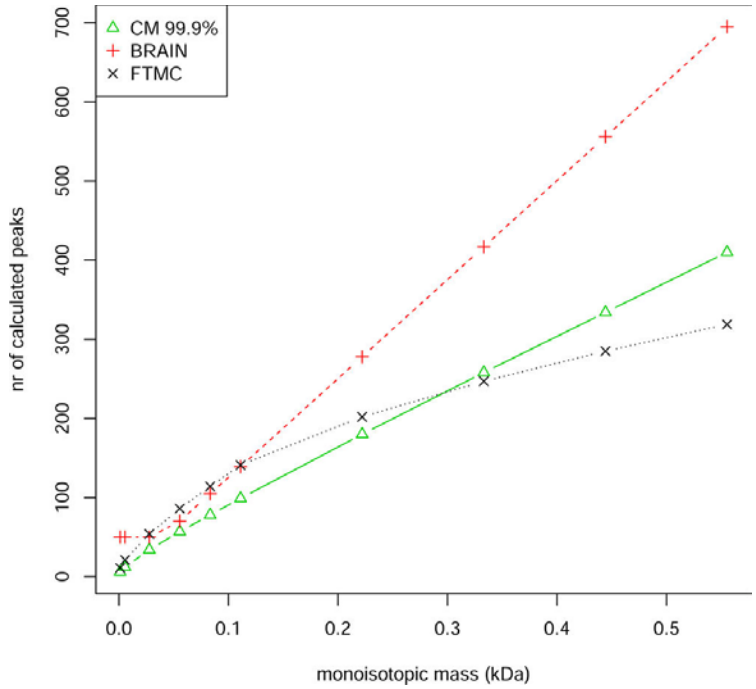


Figure 18 Number of peaks acquired for different heuristics. CM 99.9% (triangle), 99.9% coverage of the relative abundance; BRAIN (plus sign), heuristic based on (Eq.20) ; FTMC (cross sign), heuristic used by FTMC. Note that all heuristics yield at least 99.9% coverage.

The main limitation or feature of the BRAIN method is that it requires the calculation starting from the lightest isotope variant. As a comparison, the FTMC method can choose which region in the isotopic distribution to be computed. However, the feature (or limitation) of BRAIN allows flexible control of the stopping criteria, *e.g.* a 99.9% cumulative abundance. This is especially convenient for small molecules or fragments, where only a very small number of peaks (3 or 4) need to be calculated.

In order to evaluate the contribution of programming language, 57,930 proteins tested in the BRAIN application note (119) were processed using BRAIN in C++ (v0.9.6)

and in R (v1.6.4). Both calculations were conducted on a Core i7 870 2.93GHz (4GB RAM) machine. The required number of peaks were set corresponding to heuristic (Eq.20). It took 80s for the C++ program to produce the output file, while the corresponding R script took 15.5 min.

Furthermore, 10 AVERAGINE molecules described in the FTMC method (123) were used for comparing the performance of BRAIN in C++ and FTMC. In order to remove the interruption of heuristics, we ran both programs for each molecule separately using the default heuristics described in (123), and repeated this procedure 100 times. We then took the minimum time from 100 runs in order to minimize inevitable side effects such as garbage collection and interrupts. It should be noted that the minimum time still includes the initialization operations for the program, such as reading the input file, writing the result file and loading user-defined parameters. These operations will not change the theoretical asymptotic complexity but can obscure the timing performance when not appropriately accounted for, as seen in Figure 19. The result suggests that the performance of BRAIN in C++ is very stable, while there is tiny fluctuation of time for FTMC. It also shows that BRAIN in C++ consumed less time than FTMC for the same molecule, which may be attributed to the slight performance boost of C++ language.

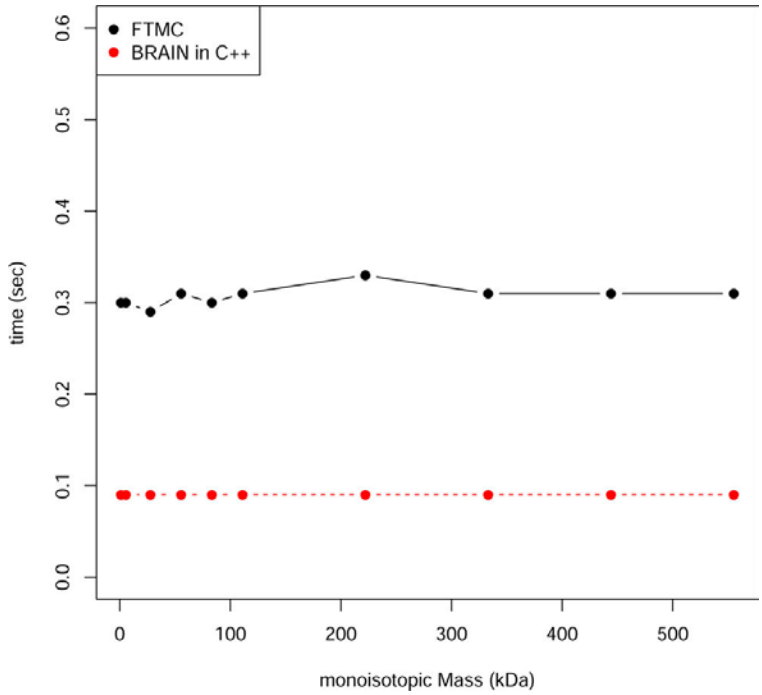


Figure 19 Average elapsed system time for FTMC and BRAIN on each AVERAGINE molecules. For each molecule, the programs were called 100 times separately and the minimum time was recorded.

For intensive batch processing, algorithm optimization is an important factor affecting the performance of an algorithm. Recurrent variables can be computed in advance and stored into the memory for reuse. For the C++ implementation of the BRAIN method, the roots and their exponents were pre-calculated. To compare the efficiency of batch processing, we run both programs for each molecule in the AVERAGINE dataset separately using the default heuristics and specify the molecule 100 times in the input file. For each molecule a result file is written as output. We argue that this comparison is more transparent than using the '–r100' option in FTMC. The obtained elapsed system time was divided by 100 and presented in Figure 20. Indeed, we

can see that the preloading operations have an impact on the timing procedure. When the monoisotopic mass was below 200 kDa, the BRAIN method with heuristic (Eq.20) ran faster than the FTMC method, regardless of its inferiority in time complexity. As expected, the BRAIN method with heuristic (Eq.20) displayed a quadratic trend in function of the molecular mass (red dots). Note that the BRAIN heuristic (Eq.20) has a linear relation between the number of peaks and the mass of a molecule, as shown in Figure 18. Figure 20 also displays a timing trend for FTMC that is proportional to the square root of the mass as presented in Figure 18 by Fernandez-de-Cossio Diaz and Fernandez-de-Cossio1. As we stated above, the time trend difference between the two algorithms is mainly caused by the difference of heuristics they use. When the factor is eliminated, a linear trend shows up for both algorithms (Figure 19). Once again, both BRAIN with the heuristic (Eq.20) and FTMC overestimate the number of required peaks in order to achieve a 99.9% coverage (green dot), which suggests the potential for improvement for both methods.

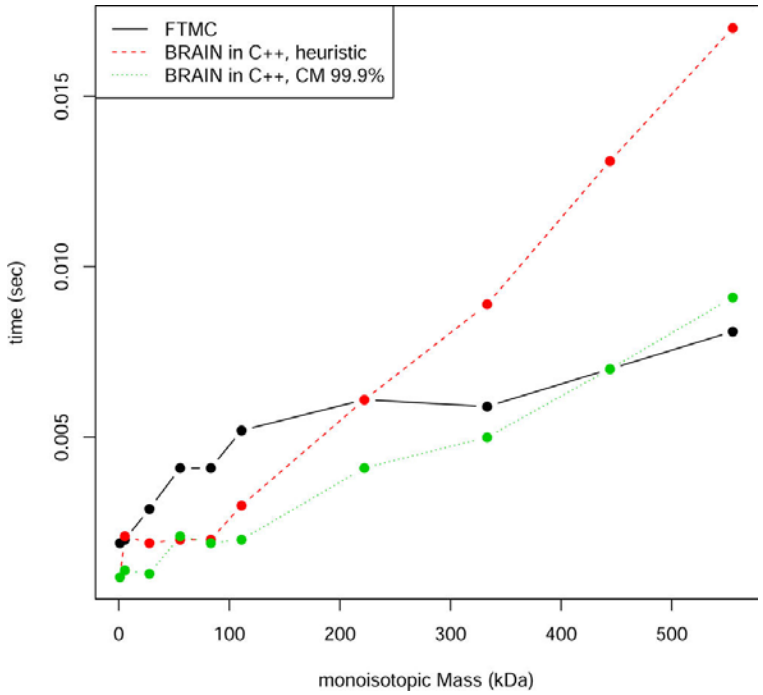


Figure 20 Average system time elapsed for FTMC and BRAIN with internal parameters preloaded. The input file for each program contains the same molecules repeating 100 times.

To sum, the C++ implementation of BRAIN (119) is particularly useful, not only because of the favorable speed but also because of the ease of integration into automatic pipeline analysis. The module works as an independent program, and was also integrated as part of our in-house deconvolution/deisotoping program SimpleFinder. The source code and binaries of the C++ implementation of BRAIN are available at https://code.google.com/p/brain-isotopic-distribution/.

5.2.3 Workflow for Identifying Monoisotopic Peaks

An in-house deconvolution/deisotoping program SimpleFinder was designed to assist in the automatic HS sequencing using HS-SEQ (Figure 21). The whole process

consists of envelop detection stage and envelop optimization stage, as discussed by Liu *et. al.* (120), with harmonics (artifact product ions) in FTMS treated.

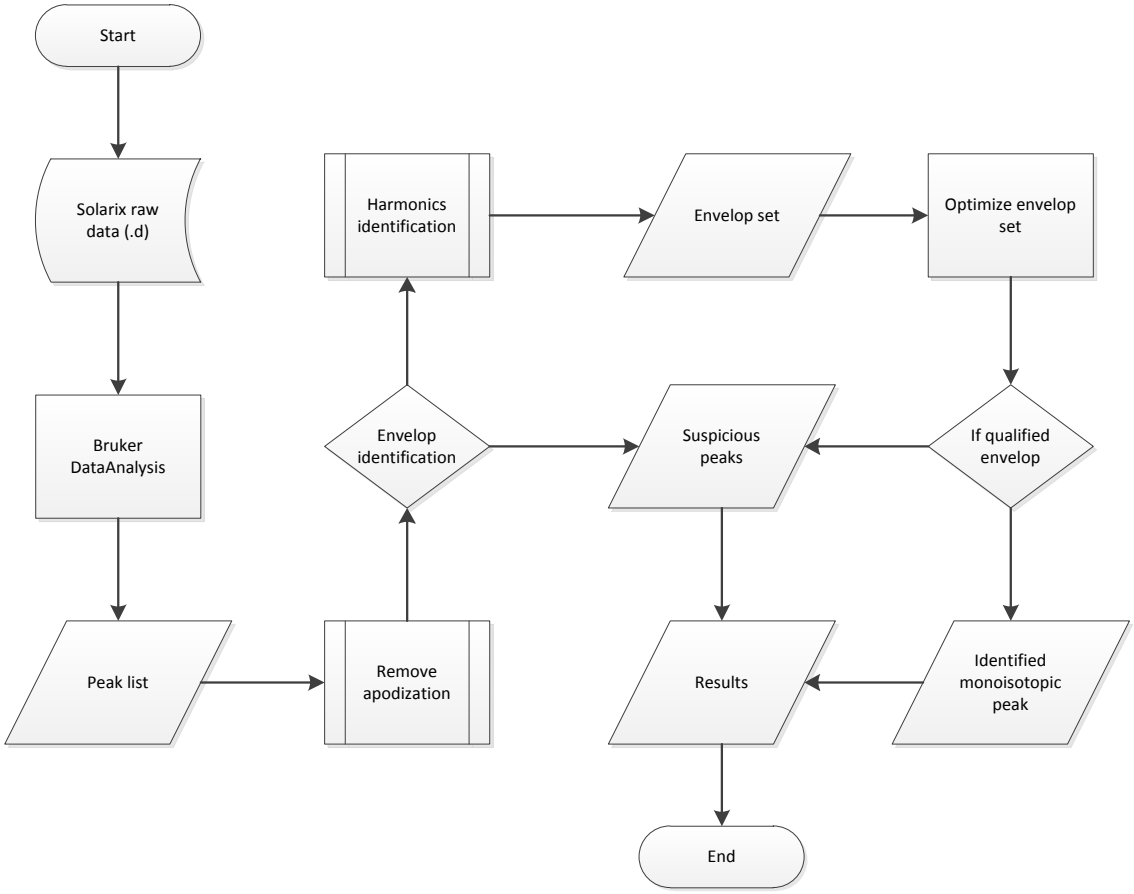


Figure 21 Flowchart of deconvolution/deisotoping in the SimpleFinder program. The workflow consists of envelop (isotopic cluster) identification and envelop optimization step to identify potential isotopic clusters and therefore their corresponding monoisotopic peaks.

In the envelop detection stage, the program iterated over the peak list; it treated each peak it met as a potential monoisotopic peak, enumerated all possible charge states

to find candidate envelops, and explored the sulfur numbers κ to optimize the fitting between the theoretical isotopic distribution (121) and the candidate envelops. The *S/N* threshold was set to 10 for the most abundant peak (also the monoisotopic peak in HS tandem mass spectra) in a candidate envelope, and 5 for the rest of the isotope peaks.

In the envelop optimization stage, the program detects abnormal isotopic peaks which has intensity either above or below given theoretical threshold, *e.g.* 30%, and makes hypothesis of the existence of an partial overlapping isotopic cluster. The process is repeated until all abnormal peaks are resolved, or the hypothetic isotopic cluster is rejected once the inclusion of it fails to improve the overall fitting score.

The output of SimpleFinder includes two files: one is a list of monoisotopic peaks with identified charge states, and the other contains a list of suspicious monoisotopic peaks with no charge information due to their low abundances and incomplete isotopic cluster patterns. HS-SEQ currently only took the former file for consideration.

5.3 Data Visualization

Data visualization is an important step to demonstrate the internal structure of data sets, support decision-making as well as hypothesis validation.

In the context of mass spectrometry data, one of the most demanding tasks in visualization is to demonstrate interesting features (*e.g.* peaks in the spectrum) and associate peaks with structural annotations (either composition or fragment information). Identification of molecule sequence may further require the connecting between peak annotations in tandem MS and candidate sequence. Scores and meta data information

associated with MS and tandem MS may also be included to better support user decision. Almost all common software suites in mass spectrometry analysis provide visualization function. MZmine 2 (126) supports visualization of chromatogram plot and spectrum plot. Users are able to view the data in either 2D or 3D view, and manually check the peak-picking performance. mMass (127) provides gel view (for comparing spectra from different samples), normalized view (for comparing spectra from different machines), spectrum flipping (general comparison between two spectra), spectrum ruler and label tools. Peptagram (128) uses HTML5 Canvas to visualize identified sequence of a single proteomics experiment, and provide graphical comparison of multiple experiments.

Visualization of mass spectrometry data typically requires participation of users. Users' expertise in instrument and spectrum interpretation is indispensable for guiding the data analysis. A full-fledged visualization tool should be able to interact closely with user, on one hand, by providing users with supportive information for decision making, and on the other hand, by taking users' input and update the inferred results.

In order to support automatic annotation of tandem mass spectra, I designed a JavaScript tool, SpectrumAnnotation, which automatically labels interesting product ions and allows user's operation on the spectrum. The prototype algorithm was implemented in R (Figure 22) and the user interface was implemented by JavaScript D3 library (129) for data-driven visualization and HTML5 Canvas for fast image rendering. The source code of SpectrumAnnotation is available from https://github.com/lamarck2008/SpectraAnnotation.

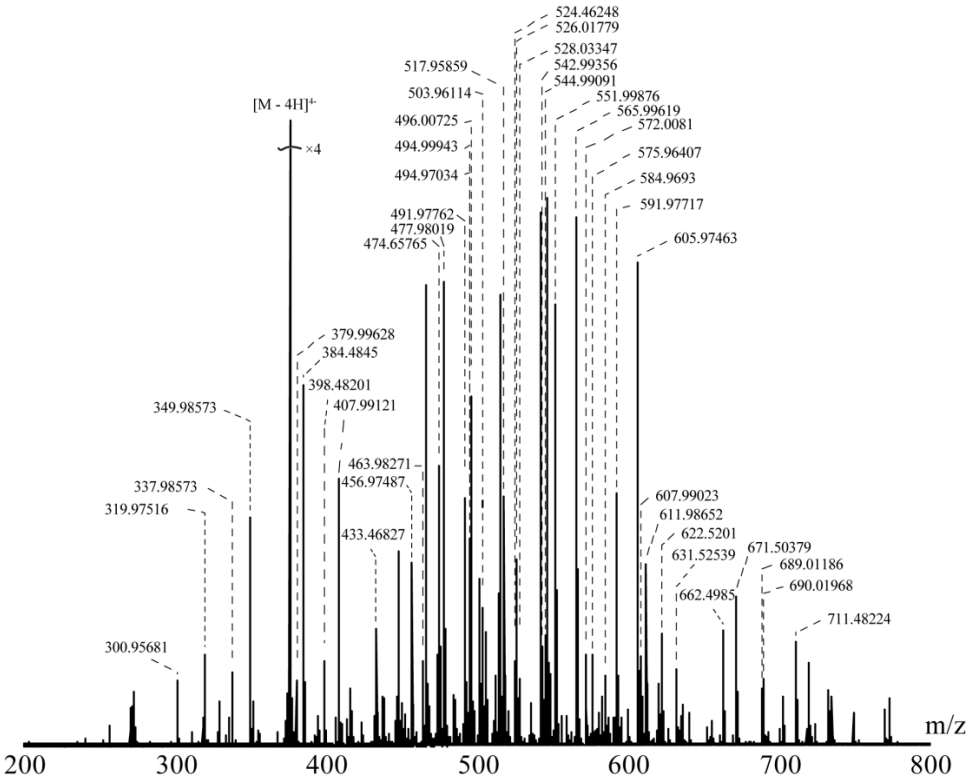


Figure 22 SpectrumAnnotation for automatic peak labeling. The relative positions of the labels in the spectra are inferred by the program based on neighboring label positions. The prototype was implemented in R and the image was modified in Illustrator.

SpectrumAnnotation was implemented in a focus+context style (Figure 23), where the "context" region provides a profile of the data and the "focus" region provides a zoomed-in picture of the spectrum based on user's selection on the "context" canvas. All the user-interactive features such as zoom-in, zoom out, drag and pan, were implemented by d3.js library and Scalable Vector Graphics (SVG) technique. The rendering speed of SVG plot largely depends on the number of graphic objects (*e.g.* circles, lines), and plunges dramatically when the object number is at thousands. This performance does not meet the requirement of whole-spectrum visualization, which

usually contains millions of data points. In order to improve the rendering performance, I designed a hybrid visualization method that combines the flexibility of d3 in controlling graphic objects and rendering performance of HTML5 Canvas.

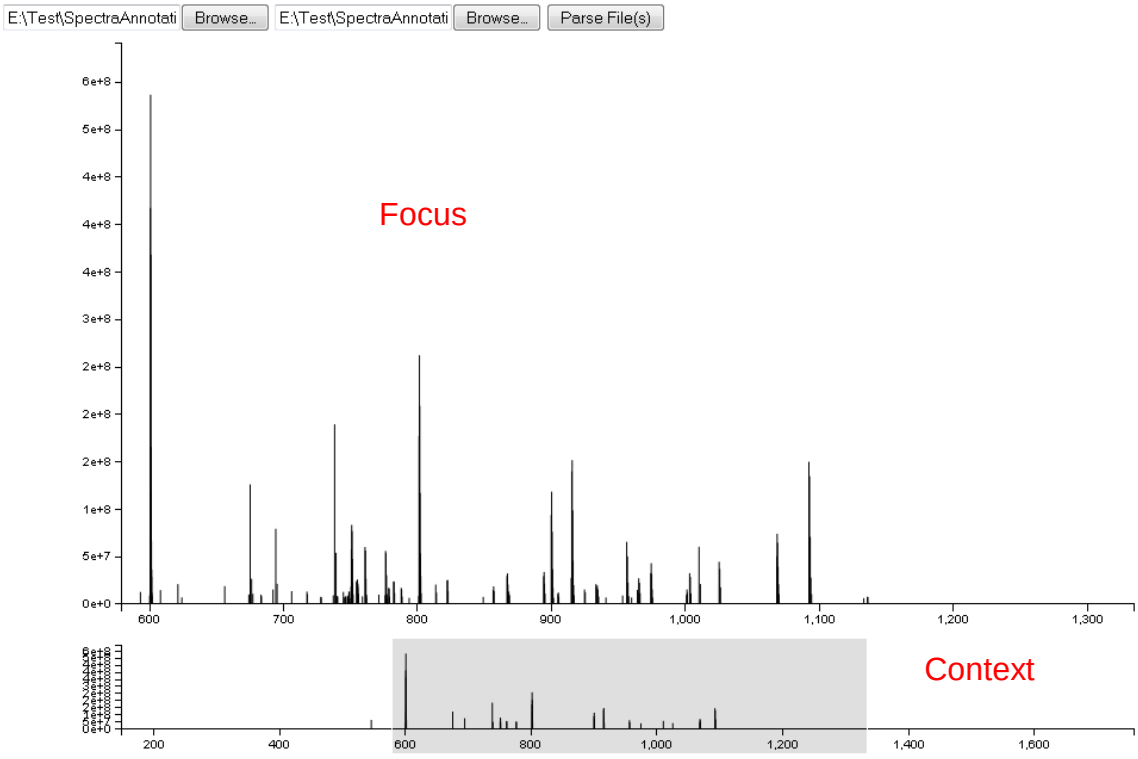


Figure 23 SpectrumAnnotation in "focus + context" style. A dataset with 203,967 peaks is used to demonstrate the layout of SpectrumAnnotation.

Since SpectrumAnnotation requires no background of the precursor sequence, it can be applied to general annotation of tandem mass spectra once the spectrum list and peak assignment list are given. Further work will focus on the integration of SpectrumAnnotation with HS-SEQ to support automatic HS sequencing. Users will be

able to tweak the parameters or update the peak annotations directly from web interface, and the corresponding sequencing results will be generated simultaneously.

Chapter 6 Conclusion and Future Directions

6.1 Summary

This thesis is on the methodology of identifying heparan sulfate fine structure using high-resolution tandem mass spectrometry, and more generally, the methodology of interpreting tandem mass spectra containing spurious peaks. Due to the paucity of comprehensive discussion on factors causing mis-interpretation and deficiency of current methods in differentiating HS isomers, this thesis focuses on formalizing the ambiguity problem in peak annotation, and designing strategies to integrate the ambiguous information for sequencing.

Chapter 2 reviews current identification algorithms in proteomics, and discusses progresses in glycomics and glycoproteomics. One hallmark of tandem MS-based sequencing algorithms is that it is much easier to implement in from sequence to fragments rather than vice versa. It's no wonder that database search methods remain the standard identification procedure for the last two decades. Algorithms in this category try to integrate knowledge of instrument, fragmentation methods and fragmentation patterns in order to effectively reduce the search space. Since mass spectrometry technique is advancing so fast, researchers have to reinvent the wheel frequently in order to cater to the new rules (either chemical or instrumental). In contrast, *de novo* sequencing algorithms try to build up the relationship between peaks, which allows rapid identification of sequence tags. Game changes when the spectra are full of ambiguous peaks: database search methods may result in a high scoring threshold for confident identification, and *de novo* sequencing methods may detour to suboptimal sequences

when random matches of mass differences frequently occur. Even the optimal path does not guarantee the correct identification (23). Therefore, a working algorithm based on ambiguous peak annotations cannot simply follow precedent work in related fields and requires tailor-made design.

Chapter 3 overviews HS-SEQ, the first algorithm for identifying heparan sulfate fine structure using high resolution tandem mass spectrometry. Compared with naive methods, HS-SEQ excels in both sensitivity and specificity, and usually takes seconds to generate sequencing results. The results of HS-SEQ also reflects the spectrum quality, which suggests the potential to design a quality control (QC) method based on the reports from HS-SEQ. The fundamental difference between HS-SEQ and other algorithms in related fields is that HS-SEQ focuses on inferring the most confident and informational peaks instead of global optimization of the sequence. With a full set of advantages in sequencing, HS-SEQ is beneficial to confirming synthesized HS structures. We also compared results from Bruker Solarix NETD with Thermo Orbitrap NETD (data not shown). Although the two have different fragmentation preferences, their respective results agreed with each other. This also suggests that HS-SEQ is in principle an instrument-independent approach for sequencing.

Chapter 4 discusses MULTI-HS-SEQ (or HS-SEQ+), an error-tolerant algorithm extended from HS-SEQ. Different from HS-SEQ, which tries to pick out the currently most confident peaks for improving the sequencing results, MULTI-HS-SEQ works with complicated situations where there might be conflicting peak interpretations with equal confidence values. Integrating conflicting peak interpretations will unavoidably

complicate the graph model and causes multiple sequences to be identified at the same time. The main feature of MULTI-HS-SEQ is that it tries to minimize the number of top-layer candidate sequences while maintain the advantages of HS-SEQ in specificity, sensitivity and running time. It is still too hasty to conclude that MULTI-HS-SEQ can solve the mixture sequencing problem, but it does take the co-fragmentation of mixture into consideration. This situation is not rare when the sample is from natural source and highly heterogeneous. The feature of MULTI-HS-SEQ in tolerating conflicting sulfation information will make it suitable in exploring unknown sulfation patterns of HS.

Chapter 5 introduces the architecture of HS sequencing pipeline, and focuses on developing modules assisting the pre-processing and post-processing of HS data: a new deconvolution/deisotoping module SimpleFinder designed specifically for identifying the HS fragments ions with large variation of sulphur content, and a visualization tool *SpectrumAnnotation* for automatic interpretation of the spectrum. These modules remove the need of tedious and error-prone human curation procedure and can effectively improve the sequencing throughput. Further study will focus on optimizing the preprocessing algorithms, and user-interactive functionality in spectrum visualization.

6.2 Generalization to Sequencing of Other Molecules Using Tandem Mass Spectrometry

As the development of instrument and fragmentation methods, the ambiguity of peak annotations and “chimeric spectra” (tandem MS with co-fragmentation) are becoming the remaining reasons causing an identification algorithm to fail. Ideally, if we

can unequivocally assign all fragments (assuming only one sequence fragmented), we are able to recover the original sequence accurately and rapidly with linear time complexity. An optimized program should be able to finish this within a second. In this thesis, HS-SEQ serves to exemplify this basic idea.

Peptide identification is dominated by database search methods, which usually require high-confidence protein database, clear understanding of fragmentation preference, elaborately designed scoring function and solid FDR calculation. Any deficiency in each of the process may significantly affect the final performance. Sequences with specific features (*e.g.* splicing variant, single amino acid mutation, symmetric structure, and rare PTMs) require tailored identification strategy. Assumptions or conditions may be implicitly nullified when one attempts to apply the classical strategies to new types of molecules (*e.g.* applying FDR in proteomics to glycoproteomics). The sequencing of peptide is essentially arranging a set of residues given their total composition (although determining the correct composition sometimes can still be difficult). Glycan sequencing takes linkage information into account, which can be addressed separately once the residue order is determined. A robust, efficient and instrument-independent sequencing algorithm is possible if one can successfully improve the confidence of peak annotations (either through labeling or computational inference), and design a proper framework to integrate these annotations. For sequence with large modification moiety (*e.g.* glycopeptide), it is essentially no different from a sequence with missing fragments. Simple brute-force method such as bloom filter (130) may help

look up the composition of the missing part efficiently. Fragmentation within the moiety can further clarify its internal organization.

6.3 Application to Studying Heparan Sulfate Proteoglycan and Protein Interaction

Understanding the binding mechanism between HS and protein is important to guide the development of new drugs, as HS mediates many growth factor signaling pathways and tunes their activities. The binding largely depends on the modification pattern (sulfation, epimerization) and domain organization of HS chains. HS-SEQ and its successor MULTI-HS-SEQ (HS-SEQ+) provide practical solutions to identify the fine structures of HS and hopefully other GAG molecules. Although HS-SEQ currently doesn't assume the difference between IdoA and GlcA, this information may be inferred from biosynthetic rules as well as diagnostic ions. Development of HS library may also contribute to understanding the distribution of epimerization and its biological significance.

To date, a few heparin-derived drugs have been developed. In addition to the well-known roles of heparin in anticoagulation, HS mimetics such as PI-88 and PG 545 can also mediate the inhibition of heparanase activities, and therefore interfere heparanase-induced HSPG degradation during metastasis. Besides, these mimetics can bind to growth factor VEGF and FGF-2, and inhibiting the growth factors' activities in stimulating tumor angiogenesis.

Compared with the wide involvement of HS in mediating biological functions in different animal organ systems, the development of HS-derived drugs is still in its

infancy (131). This can be attributed to the complexity of binding mechanisms between HS chains and HSBPs, and further to the lack of clearly resolved HS sequences. Considering the advance of chemical and chemoenzymatic synthesis of HS oligosaccharides, new fragmentation methods to maintain the sulfate intactness of HS, and computational methods such as HS-SEQ (106) to accurately identify the sequence, the stale situation will be dramatically changed in the next few years. We expect to see more HS-derived drugs into clinical trial in the near future.

BIBLIOGRAPHY

1. Medeiros, G. F., Mendes, A., Castro, R. A. B., Baú, E. C., Nader, H. B., and Dietrich, C. P. (2000) Distribution of sulfated glycosaminoglycans in the animal kingdom: widespread occurrence of heparin-like compounds in invertebrates. *Biochimica et Biophysica Acta (BBA) - General Subjects* 1475, 287–294

2. Guerardel, Y., Czeszak, X., Sumanovski, L. T., Karamanos, Y., Popescu, O., Strecker, G., and Misevic, G. N. (2004) Molecular Fingerprinting of Carbohydrate Structure Phenotypes of Three Porifera Proteoglycan-like Glyconectins. *Journal of Biological Chemistry* 279, 15591–15603

3. Bishop, J. R., Schuksz, M., and Esko, J. D. (2007) Heparan sulphate proteoglycans fine-tune mammalian physiology. *Nature* 446, 1030–1037

4. Parish, C. R. (2006) The role of heparan sulphate in inflammation. *Nature Reviews Immunology* 6, 633–643

5. Sasisekharan, R., Shriver, Z., Venkataraman, G., and Narayanasami, U. (2002) Roles of heparan-sulphate glycosaminoglycans in cancer. *Nature Reviews Cancer* 2, 521–528

6. Knelson, E. H., Nee, J. C., and Blobe, G. C. (2014) Heparan sulfate signaling in cancer. *Trends in Biochemical Sciences* 39, 277–288

7. Makarenkova, H. P., Hoffman, M. P., Beenken, A., Eliseenkova, A. V., Meech, R., Tsau, C., Patel, V. N., Lang, R. A., and Mohammadi, M. (2009) Differential Interactions of FGFs with Heparan Sulfate Control Gradient Formation and Branching Morphogenesis. *Science Signaling* 2, ra55–ra55

8. Coles, C. H., Shen, Y., Tenney, A. P., Siebold, C., Sutton, G. C., Lu, W., Gallagher, J. T., Jones, E. Y., Flanagan, J. G., and Aricescu, A. R. (2011) Proteoglycan-Specific Molecular Switch for RPTPσ Clustering and Neuronal Extension. *Science* 332, 484–488

9. Xu, D., and Esko, J. D. (2014) Demystifying Heparan Sulfate–Protein Interactions. *Annual Review of Biochemistry* 83, 129–157

10. Kreuger, J., Spillmann, D., Li, J., and Lindahl, U. (2006) Interactions between heparan sulfate and proteins: the concept of specificity. *The Journal of Cell Biology* 174, 323–327

11. Kreuger, J., and Kjellén, L. (2012) Heparan Sulfate Biosynthesis Regulation and Variability. *Journal of Histochemistry & Cytochemistry* 60, 898–907

12. Ueno, M., Yamada, S., Zako, M., Bernfield, M., and Sugahara, K. (2001) Structural Characterization of Heparan Sulfate and Chondroitin Sulfate of Syndecan-1 Purified from Normal Murine Mammary Gland Epithelial Cells COMMON PHOSPHORYLATION OF XYLOSE AND DIFFERENTIAL SULFATION OF GALACTOSE IN THE PROTEIN LINKAGE REGION TETRASACCHARIDE SEQUENCE. *Journal of Biological Chemistry* 276, 29134–29140

13. Sarrazin, S., Lamanna, W. C., and Esko, J. D. (2011) Heparan Sulfate Proteoglycans. *Cold Spring Harbor Perspectives in Biology* 3, a004952

14. Lindahl, U., and Li, J. (2009) Interactions between heparan sulfate and proteins—design and functional implications. *International review of cell and molecular biology* 276, 105–159

15. Olson, S. T., Halvorson, H. R., and Björk, I. (1991) Quantitative characterization of the thrombin-heparin interaction. Discrimination between specific and nonspecific binding models. *Journal of Biological Chemistry* 266, 6342–6352

16. Jones, C. J., and Larive, C. K. (2011) Carbohydrates: Cracking the glycan sequence code. *Nature Chemical Biology* 7, 758–759

17. Ly, M., Iii, F. E. L., Laremore, T. N., Toida, T., Amster, I. J., and Linhardt, R. J. (2011) The proteoglycan bikunin has a defined sequence. *Nature Chemical Biology* 7, 827–833

18. Zhao, X., Yang, B., Solakylidirim, K., Joo, E. J., Toida, T., Higashi, K., Linhardt, R. J., and Li, L. (2013) Sequence Analysis and Domain Motifs in the Porcine Skin Decorin Glycosaminoglycan Chain. *Journal of Biological Chemistry* 288, 9226–9237

19. Thacker, B. E., Xu, D., Lawrence, R., and Esko, J. D. (2014) Heparan sulfate 3-O-sulfation: A rare modification in search of a function. *Matrix Biology* 35, 60–72

20. Li, L., Ly, M., and Linhardt, R. J. (2012) Proteoglycan sequence. *Molecular BioSystems* 8, 1613

21. Zaia, J. (2013) Glycosaminoglycan Glycomics Using Mass Spectrometry. *Molecular & Cellular Proteomics* 12, 885–892

22. Wolff, J. J., Chi, L., Linhardt, R. J., and Amster, I. J. (2007) Distinguishing Glucuronic from Iduronic Acid in Glycosaminoglycan Tetrasaccharides by Using Electron Detachment Dissociation. *Analytical Chemistry* 79, 2015–2022

23. Allmer, J. (2011) Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Review of Proteomics* 8, 645–657

24. Medzihradszky, K. F., and Chalkley, R. J. (2013) Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass Spectrometry Reviews*, n/a–n/a

25. Mann, M., and Wilm, M. (1994) Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry* 66, 4390–4399

26. Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. (1999) De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology* 6, 327–342

27. Li, Y. F., and Radivojac, P. (2012) Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics* 13, S4

28. Matthiesen, R., Trelle, M. B., Højrup, P., Bunkenborg, J., and Jensen, O. N. (2005) VEMS 3.0: Algorithms and Computational Tools for Tandem Mass Spectrometry Based Identification of Post-translational Modifications in Proteins. *Journal of Proteome Research* 4, 2338–2347

29. Shan, B., Ma, B., Zhang, K., and Lajoie, G. (2008) Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *Journal of Bioinformatics and Computational Biology* 06, 77–91

30. Apte, A., and Meitei, N. S. (2010) in *Functional Glycomics*, Methods in Molecular Biology., ed Li J (Humana Press), pp 269–281.

31. Tang, H., Mechref, Y., and Novotny, M. V. (2005) Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* 21, i431–i439

32. Wang, J., Bourne, P. E., and Bandeira, N. (2011) Peptide Identification by Database Search of Mixture Tandem Mass Spectra. *Molecular & Cellular Proteomics* 10, M111.010017

33. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5, 976–989

34. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567

35. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466–1467

36. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open Mass Spectrometry Search Algorithm. *Journal of Proteome Research* 3, 958–964

37. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis. *Journal of Proteome Research* 6, 654–661

38. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of Accurate Mass Measurement (±10 ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching. *Analytical Chemistry* 71, 2871–2882

39. Chalkley, R. J., Baker, P. R., Medzihradszky, K. F., Lynn, A. J., and Burlingame, A. L. (2008) In-depth Analysis of Tandem Mass Spectrometry Data from Disparate Instrument Types. *Molecular & Cellular Proteomics* 7, 2386–2398

40. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* 10, 1794–1805

41. Wenger, C. D., and Coon, J. J. (2013) A Proteomics Search Algorithm Specifically Designed for High-Resolution Tandem Mass Spectra. *Journal of Proteome Research* 12, 1377–1386

42. Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013) Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS* 13, 22–24

43. Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., and Mechtler, K. (2014) MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. *Journal of Proteome Research* 13, 3679–3684

44. Sadygov, R. G., Cociorva, D., and Yates, J. R. (2004) Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nature Methods* 1, 195–202

45. Steen, H., and Mann, M. (2004) The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology* 5, 699–711

46. Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J.,

George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H., and Pandey, A. (2014) A draft map of the human proteome. *Nature* 509, 575–581

47. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F., and Kuster, B. (2014) Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587

48. Ma, B., and Johnson, R. (2012) De Novo Sequencing and Homology Searching. *Molecular & Cellular Proteomics* 11, O111.014902

49. Pearson, W. R., and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences* 85, 2444–2448

50. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., and others (2005) The universal protein resource (UniProt). *Nucleic acids research* 33, D154–D159

51. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35, D61–D65

52. Choudhary, J. S., Blackstock, W. P., Creasy, D. M., and Cottrell, J. S. (2001) Matching peptide mass spectra to EST and genomic DNA databases. *Trends in Biotechnology* 19, Supplement 1, 17–22

53. Na, S., Jeong, J., Park, H., Lee, K.-J., and Paek, E. (2008) Unrestrictive Identification of Multiple Post-translational Modifications from Tandem Mass Spectrometry Using an Error-tolerant Algorithm Based on an Extended Sequence Tag Approach. *Molecular & Cellular Proteomics* 7, 2452–2463

54. Ahrné, E., Müller, M., and Lisacek, F. (2010) Unrestricted identification of modified proteins using MS/MS. *PROTEOMICS* 10, 671–686

55. Matthiesen, R., Azevedo, L., Amorim, A., and Carvalho, A. S. (2011) Discussion on common data analysis strategies used in MS-based proteomics. *PROTEOMICS* 11, 604–619

56. Allmer, J. (2012) A Call for Benchmark Data in Mass Spectrometry-Based Proteomics. *Journal of Integrated OMICS* 2,

57. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* 4, 207–214

58. Gupta, N., Bandeira, N., Keich, U., and Pevzner, P. A. (2011) Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong. *Journal of The American Society for Mass Spectrometry* 22, 1111–1120

59. Kwon, T., Choi, H., Vogel, C., Nesvizhskii, A. I., and Marcotte, E. M. (2011) MSblender: A Probabilistic Approach for Integrating Peptide Identifications from Multiple Database Search Engines. *Journal of Proteome Research* 10, 2949–2958

60. Edwards, N., Wu, X., and Tseng, C.-W. (2009) An Unsupervised, Model-Free, Machine-Learning Combiner for Peptide Identifications from Tandem Mass Spectra. *Clinical Proteomics* 5, 23–36

61. Nahnsen, S., Bertsch, A., Rahnenführer, J., Nordheim, A., and Kohlbacher, O. (2011) Probabilistic Consensus Scoring Improves Tandem Mass Spectrometry Peptide Identification. *Journal of Proteome Research* 10, 3332–3343

62. Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008) OpenMS – An open-source software framework for mass spectrometry. *BMC Bioinformatics* 9, 163

63. Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, A. I. (2011) iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Molecular & Cellular Proteomics* 10, M111.007690

64. Barsnes, H., Vaudel, M., Colaert, N., Helsens, K., Sickmann, A., Berven, F. S., and Martens, L. (2011) compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics* 12, 70

65. Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., and Martens, L. (2011) SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 11, 996–999

66. Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L., and Deutsch, E. W. (2013) Combining Results of Multiple Search Engines in Proteomics. *Molecular & Cellular Proteomics* 12, 2383–2393

67. Guthals, A., and Bandeira, N. (2012) Peptide Identification by Tandem Mass Spectrometry with Alternate Fragmentation Modes. *Molecular & Cellular Proteomics* 11, 550–557

68. Na, S., and Paek, E. (2014) Software eyes for protein post-translational modifications. *Mass Spectrometry Reviews*, n/a–n/a

69. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* 17, 2337–2342

70. Heredia-Langner, A., Cannon, W. R., Jarman, K. D., and Jarman, K. H. (2004) Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data. *Bioinformatics* 20, 2296–2304

71. Chen, T., Kao, M.-Y., Tepel, M., Rush, J., and Church, G. M. (2001) A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology* 8, 325–337

72. Mo, L., Dutta, D., Wan, Y., and Chen, T. (2007) MSNovo: A Dynamic Programming Algorithm for de Novo Peptide Sequencing via Tandem Mass Spectrometry. *Analytical Chemistry* 79, 4870–4878

73. Lu, B., and Chen, T. (2003) A Suboptimal Algorithm for De Novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology* 10, 1–12

74. Chi, H., Chen, H., He, K., Wu, L., Yang, B., Sun, R.-X., Liu, J., Zeng, W.-F., Song, C.-Q., He, S.-M., and Dong, M.-Q. (2013) pNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra. *Journal of Proteome Research* 12, 615–625

75. Zhang, Z. (2004) De Novo Peptide Sequencing Based on a Divide-and-Conquer Algorithm and Peptide Tandem Spectrum Simulation. *Analytical Chemistry* 76, 6374–6383

76. Spengler, B. (2004) De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry. *Journal of the American Society for Mass Spectrometry* 15, 703–714

77. Olson, M. T., Epstein, J. A., and Yergey, A. L. (2006) De Novo Peptide Sequencing Using Exhaustive Enumeration of Peptide Composition. *Journal of the American Society for Mass Spectrometry* 17, 1041–1049

78. Jeong, K., Kim, S., and Pevzner, P. A. (2013) UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics* 29, 1953–1962

79. Liu, X., Dekker, L. J. M., Wu, S., Vanduijn, M. M., Luider, T. M., Tolić, N., Kou, Q., Dvorkin, M., Alexandrova, S., Vyatkina, K., Paša-Tolić, L., and Pevzner, P. A. (2014) De Novo Protein Sequencing by Combining Top-Down and Bottom-Up Tandem Mass Spectra. *Journal of Proteome Research* 13, 3241–3248

80. Creasy, D. M., and Cottrell, J. S. (2004) Unimod: Protein modifications for mass spectrometry. *PROTEOMICS* 4, 1534–1536

81. Garavelli, J. S. (2004) The RESID Database of Protein Modifications as a resource and annotation tool. *PROTEOMICS* 4, 1527–1533

82. Anderson, N. L., and Anderson, N. G. (2002) The Human Plasma Proteome History, Character, and Diagnostic Prospects. *Molecular & Cellular Proteomics* 1, 845–867

83. Phanstiel, D., Brumbaugh, J., Berggren, W. T., Conard, K., Feng, X., Levenstein, M. E., McAlister, G. C., Thomson, J. A., and Coon, J. J. (2008) Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. *Proceedings of the National Academy of Sciences* 105, 4093–4098

84. Pevzner, P. A., Mulyukov, Z., Dancik, V., and Tang, C. L. (2001) Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry. *Genome Research* 11, 290–299

85. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nature Biotechnology* 23, 1562–1567

86. Searle, B. C., Dasari, S., Wilmarth, P. A., Turner, M., Reddy, A. P., David, L. L., and Nagalla, S. R. (2005) Identification of Protein Modifications Using MS/MS de Novo Sequencing and the OpenSea Alignment Algorithm. *Journal of Proteome Research* 4, 546–554

87. HAN, Y., MA, B., and ZHANG, K. (2005) SPIDER: SOFTWARE FOR PROTEIN IDENTIFICATION FROM SEQUENCE TAGS WITH DE NOVO SEQUENCING ERROR. *Journal of Bioinformatics and Computational Biology* 03, 697–716

88. Na, S., Bandeira, N., and Paek, E. (2012) Fast Multi-blind Modification Search through Tandem Mass Spectrometry. *Molecular & Cellular Proteomics* 11, M111.010199

89. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2006) ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. *Molecular & Cellular Proteomics* 5, 935–948

90. Fu, Y., Xiu, L.-Y., Jia, W., Ye, D., Sun, R.-X., Qian, X.-H., and He, S.-M. (2011) DeltAMT: A Statistical Algorithm for Fast Detection of Protein Modifications From LC-MS/MS Data. *Molecular & Cellular Proteomics* 10, M110.000455

91. Bandeira, N., Tsur, D., Frank, A., and Pevzner, P. A. (2007) Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences* 104, 6140–6145

92. Guthals, A., Watrous, J. D., Dorrestein, P. C., and Bandeira, N. (2012) The spectral networks paradigm in high throughput mass spectrometry. *Molecular BioSystems* 8, 2535

93. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology* 24, 1285–1292

94. Bailey, C. M., Sweet, S. M. M., Cunningham, D. L., Zeller, M., Heath, J. K., and Cooper, H. J. (2009) SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra. *Journal of Proteome Research* 8, 1965–1971

95. DiMaggio, P. A., Young, N. L., Baliban, R. C., Garcia, B. A., and Floudas, C. A. (2009) A Mixed Integer Linear Optimization Framework for the Identification and Quantification of Targeted Post-translational Modifications of Highly Modified Proteins Using Multiplexed Electron Transfer Dissociation Tandem Mass Spectrometry. *Molecular & Cellular Proteomics* 8, 2527–2543

96. Apweiler, R., Hermjakob, H., and Sharon, N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta (BBA) - General Subjects* 1473, 4–8

97. Khoury, G. A., Baliban, R. C., and Floudas, C. A. (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific Reports* 1,

98. Woodin, C. L., Maxon, M., and Desaire, H. (2013) Software for Automated Interpretation of Mass Spectrometry Data from Glycans and Glycopeptides. *The Analyst* 138, 2793–2803

99. Dallas, D. C., Martin, W. F., Hua, S., and German, J. B. (2013) Automated glycopeptide analysis—review of current state and future directions. *Briefings in Bioinformatics* 14, 361–374

100. Li, F., Glinskii, O. V., and Glinsky, V. V. (2013) Glycobioinformatics: Current strategies and tools for data mining in MS-based glycoproteomics. *PROTEOMICS* 13, 341–354

101. Domon, B., and Costello, C. E. (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal* 5, 397–409

102. Spencer, J. L., Bernanke, J. A., Buczek-Thomas, J. A., and Nugent, M. A. (2010) A Computational Approach for Deciphering the Organization of Glycosaminoglycans. *PLoS ONE* 5, e9389

103. Saad, O. M., and Leary, J. A. (2005) Heparin Sequencing Using Enzymatic Digestion and ESI-MSn with HOST: A Heparin/HS Oligosaccharide Sequencing Tool. *Analytical Chemistry* 77, 5902–5911

104. Tissot, B., Ceroni, A., Powell, A. K., Morris, H. R., Yates, E. A., Turnbull, J. E., Gallagher, J. T., Dell, A., and Haslam, S. M. (2008) Software Tool for the Structural Determination of Glycosaminoglycans by Mass Spectrometry. *Analytical Chemistry* 80, 9204–9212

105. Ceroni, A., Maass, K., Geyer, H., Geyer, R., Dell, A., and Haslam, S. M. (2008) GlycoWorkbench: A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans†. *Journal of Proteome Research* 7, 1650–1659

106. Hu, H., Huang, Y., Mao, Y., Yu, X., Xu, Y., Liu, J., Zong, C., Boons, G.-J., Lin, C., Xia, Y., and Zaia, J. (2014) A Computational Framework for Heparan Sulfate Sequencing Using High-resolution Tandem Mass Spectra. *Molecular & Cellular Proteomics* 13, 2490–2502

107. Huang, Y., Shi, X., Yu, X., Leymarie, N., Staples, G. O., Yin, H., Killeen, K., and Zaia, J. (2011) Improved Liquid Chromatography-MS/MS of Heparan Sulfate Oligosaccharides via Chip-Based Pulsed Makeup Flow. *Analytical Chemistry* 83, 8222–8229

108. Yu, X., Huang, Y., Lin, C., and Costello, C. E. (2012) Energy-Dependent Electron Activated Dissociation of Metal-Adducted Permethylated Oligosaccharides. *Analytical Chemistry* 84, 7487–7494

109. Yu, X., Jiang, Y., Chen, Y., Huang, Y., Costello, C. E., and Lin, C. (2013) Detailed Glycan Structural Characterization by Electronic Excitation Dissociation. *Analytical Chemistry* 85, 10017–10021

110. Xu, Y., Masuko, S., Takieddin, M., Xu, H., Liu, R., Jing, J., Mousa, S. A., Linhardt, R. J., and Liu, J. (2011) Chemoenzymatic Synthesis of Homogeneous Ultralow Molecular Weight Heparins. *Science* 334, 498–501

111. Wolff, J. J., Leach, F. E., Laremore, T. N., Kaplan, D. A., Easterling, M. L., Linhardt, R. J., and Amster, I. J. (2010) Negative Electron Transfer Dissociation of Glycosaminoglycans. *Analytical Chemistry* 82, 3460–3466

112. Leach, F. E., Wolff, J. J., Xiao, Z., Ly, M., Laremore, T. N., Arungundram, S., Al-Mafraji, K., Venot, A., Boons, G.-J., Linhardt, R. J., and Amster, I. J. (2011) Negative electron transfer dissociation Fourier transform mass spectrometry of glycosaminoglycan carbohydrates. *European Journal of Mass Spectrometry (Chichester, England)* 17, 167–176

113. Shi, X., Huang, Y., Mao, Y., Naimy, H., and Zaia, J. (2012) Tandem Mass Spectrometry of Heparan Sulfate Negative Ions: Sulfate Loss Patterns and Chemical Modification Methods for Improvement of Product Ion Profiles. *Journal of The American Society for Mass Spectrometry* 23, 1498–1511

114. Kailemia, M. J., Li, L., Ly, M., Linhardt, R. J., and Amster, I. J. (2012) Complete Mass Spectral Characterization of a Synthetic Ultralow-Molecular-Weight Heparin Using Collision-Induced Dissociation. *Analytical Chemistry* 84, 5475–5478

115. Zaia, J., and Costello, C. E. (2003) Tandem Mass Spectrometry of Sulfated Heparin-Like Glycosaminoglycan Oligosaccharides. *Analytical Chemistry* 75, 2445–2455

116. Huang, Y., Yu, X., Mao, Y., Costello, C. E., Zaia, J., and Lin, C. (2013) De Novo Sequencing of Heparan Sulfate Oligosaccharides by Electron-Activated Dissociation. *Analytical Chemistry* 85, 11979–11986

117. Huang, R., Liu, J., and Sharp, J. S. (2013) An Approach for Separation and Complete Structural Sequencing of Heparin/Heparan Sulfate-like Oligosaccharides. *Analytical Chemistry* 85, 5787–5795

118. Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved

isotopic distributions. *Journal of the American Society for Mass Spectrometry* 6, 229–233

119. Dittwald, P., Claesen, J., Burzykowski, T., Valkenborg, D., and Gambin, A. (2013) BRAIN: A Universal Tool for High-Throughput Calculations of the Isotopic Distribution for Mass Spectrometry. *Analytical Chemistry* 85, 1991–1994

120. Liu, X., Inbar, Y., Dorrestein, P. C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J. P., Bafna, V., and Pevzner, P. A. (2010) Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins. *Molecular & Cellular Proteomics* 9, 2772 –2782

121. Claesen, J., Dittwald, P., Burzykowski, T., and Valkenborg, D. (2012) An Efficient Method to Calculate the Aggregated Isotopic Distribution and Exact Center-Masses. *Journal of The American Society for Mass Spectrometry* 23, 753–763

122. Valkenborg, D., Mertens, I., Lemière, F., Witters, E., and Burzykowski, T. (2012) The isotopic distribution conundrum. *Mass Spectrometry Reviews* 31, 96–109

123. Fernandez-de-Cossio Diaz, J., and Fernandez-de-Cossio, J. (2012) Computation of Isotopic Peak Center-Mass Distribution by Fourier Transform. *Analytical Chemistry* 84, 7052–7056

124. Hu, H., Dittwald, P., Zaia, J., and Valkenborg, D. (2013) Comment on "Computation of Isotopic Peak Center-Mass Distribution by Fourier Transform." *Analytical Chemistry* 85, 12189–12192

125. Fourment, M., and Gillings, M. R. (2008) A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics* 9, 82

126. Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11, 395

127. Strohalm, M., Kavan, D., Novák, P., Volný, M., and Havlíček, V. (2010) mMass 3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data. *Analytical Chemistry* 82, 4648–4651

128. Rob, G. peptagram. *GitHub*,

129. Bostock, M., Ogievetsky, V., and Heer, J. (2011) D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 2301–2309

130. Bloom, B. H. (1970) Space/Time Trade-offs in Hash Coding with Allowable Errors. *Commun. ACM* 13, 422–426

131. Lindahl, U., and Kjellén, L. (2013) Pathophysiology of heparan sulphate: many diseases, few drugs. *Journal of Internal Medicine* 273, 555–571

CURRICULUM VITAE

Han Hu

Bioinformatics Program, Boston University, Boston MA 02215

Center for Biomedical Mass Spectrometry, 670 Albany Street, Suite 504, Boston, MA 02118

Tel: 617-378-7287 Email: hh1985@bu.edu

EDUCATION

2008 – 2015 Ph.D./M.S. in Bioinformatics, Boston University GPA 3.78
Research in computational mass spectrometry and bioinformatics
Thesis title: *De novo sequencing of heparan sulfate fine structure using high resolution tandem mass spectrometry*
Advisor: Joseph Zaia, Center for Biomedical Mass Spectrometry, Department of Chemistry, Boston University
Co-advisor: Yu (Brandon) Xia, Department of Bioengineering, Faculty of Engineering, McGill University, Montreal, Canada

2003 – 2007 B.E. in Bioengineering, Sichuan University GPA 3.51
Thesis title: Numerical analysis of gyrotactic bioconvection
Advisor: Mei Wang, School of Chemical Engineering, Sichuan University, Chengdu, Sichuan, China

PROFESSIONAL EXPERIENCE

2010 – 2015 Research Assistant, Bioinformatics Program & Center for Biomedical Mass Spectrometry, Boston University, Boston
Advisor: Prof. Joseph Zaia & Prof. Yu (Brandon) Xia

- Designed a novel deconvolution algorithm to detect the abnormal isotopic patterns caused by sulphur;
- Designed HS-SEQ, the first algorithm for sequencing heparan sulfate molecules, which provided product-grade sequencing performance and was a zero-to-one innovation in the field;
- Developed a Python module for identifying site-specific *N*-glycosylation on glycopeptides;
- Developed a preprocessing module for different vendor raw files based on ProteoWizard API;

- Developed a visualization tool for automatic spectrum labeling using D3 Javascript library;
- Collaborate with Thermo Fisher and other group to develop new NETD technique, and use my previous work as benchmark to test the performance of Thermo NETD.

2014 – 2014 Summer Intern, Pfizer Inc., Cambridge
Supervisor: Robert Yang

- Constructed quantitative proteomics pipeline for analyzing the latest human proteome datasets using OpenMS and TPP;
- Developed interactive web application for exploring RNASeq data using D3 library.

2011 – 2011 Review Assistant, *Nucleic Acids Research Database Issue, Oxford Journals*
Supervisor: Prof. Gary Benson

- Reviewed the function implementation and interface design of web servers submitted to NAR.

2009 – 2010 Research Assistant, Bioinformatics Program, Boston University, Boston
Advisor: Prof. Salomon Amar

- Identified significantly enriched pathways responsible for acute and chronic infections using microarray data.

2008 – 2008 Rotation Student, Bioinformatics Program, Boston University, Boston
Advisor: Prof. Mark Kon

- Predicted motifs of yeast genome using SVM, and compared the results with machine learning methods such as KNN, K-means and Random Forest (MATLAB and its plugin Spider).

2007 – 2008 Research Volunteer, School of Life Science, Sichuan University, Chengdu, China
Advisor: Prof. Xiao Li

- Developed web application for visualizing plant protein-protein interaction network.

2007 – 2007 Undergraduate Research Assistant, School of Chemical Engineering, Sichuan University, Chengdu, Sichuan, China
Advisor: Prof. Mei Wang

- Simulation of gyrotactic bioconvection using FLUENT and Perl
- Awarded Outstanding Thesis of the Year in the school for excellent performance.

2005 – 2005 Research Volunteer, School of Chemical Engineering, Sichuan University, Chengdu, Sichuan, China
Advisor: Prof. Lan Xianqiu
- Involved in preparation and *in vitro* release of Tamsulosin hydrochloride sustained-release tablets using high performance liquid chromatography (HPLC)

PUBLICATIONS

- **Hu, H.**, *et al.* A Computational Framework for Heparan Sulfate Sequencing Using High-resolution Tandem Mass Spectra. Mol Cell Proteomics 13, 2490-2502 (2014)

- Khatri, K., Staples, G., Leymarie, N., Leon, D, Turiak, L., Huang, Y., Yip, S., **Hu, H.**, *et al.* Confident Assignment of Site-Specific Glycosylation in Complex Glycoproteins in a Single Step. J. Proteome Res. doi:10.1021/pr500506z (2014).

- Yu, C., Lopez, C., **Hu, H.**, *et al.* A High-Throughput Method to Examine Protein-Nucleotide Interactions Identifies Targets of the Bacterial Transcriptional Regulatory Protein Fur. PLoS ONE 9, e96832 (2014).

- **Hu, H.**, Dittwald, P., Zaia, J. & Valkenborg, D. Comment on 'Computation of Isotopic Peak Center-Mass Distribution by Fourier Transform'. Anal. Chem. 85, 12189–12192 (2013).

- Maxwell, E., Tan, Y., Tan, Y, **Hu, H.**, *et al.* GlycReSoft: A Software Package for Automated Recognition of Glycans from LC/MS Data. PLoS ONE 7, e45474 (2012).

- Yu, W.-H., **Hu, H.**, Zhou, Q., Xia, Y. & Amar, S. Bioinformatics Analysis of Macrophages Exposed to Porphyromonas gingivalis: Implications in Acute vs. Chronic Infections. PLoS ONE 5, e15613 (2010).

CONFERENCE TALKS AND POSTERS

2014 **Hu. H.** *et al.* Identification of sulfation patterns for heparan sulfate mixtures using high resolution tandem mass spectrometry (Poster). 62nd ASMS Conference, Baltimore

2013 **Hu, H.** *et al.* Computational approach for predicting sulfation pattern of heparan sulfate using high resolution tandem mass spectrometry data (Presentation). 61st ASMS Conference, Minneapolis

2012 **Hu, H.** et al. Strategies for differentiating isomers from heparan sulfate fragmentation using tandem mass spectrometry (Poster). 60th ASMS Conference, Vancouver, Canada.

2011 **Hu, H.** *et al.* Interpretation of glycosaminoglycan liquid chromatography tandem mass spectrometry data from a network perspective (Poster). 11th International Workshop on Bioinformatics and Systems Biology, Berlin, Germany

2011 **Hu, H.** *et al*. Differential analysis of glycosaminoglycan tandem mass spectrometry profile data from the perspective of network topology (Poster). 59th ASMS Conference, Denver

CERTIFICATES

2014 Cloudera Certified Professional: Data Scientist (CCP: DS) written exam portion

2014 English for Foreign-Born Professionals (EFBP), The Boston Language Institute

2006 Qualification Certificate of Software Technology Proficiency, China

2005 The National Computer Rank Examination (Grade 4), China

HONORS AND AWARDS

2007 Top Graduation Thesis in in Sichuan University

2005 – 2006 Excellent Student Cadre of Sichuan University

2004 – 2005 Excellent Student of School of Chemical Engineering

2004 – 2005 Excellence Scholarship

2003 – 2004 Excellence Scholarship

2003 – 2004 Golden Land Scholarship

2002 The Second Prize of the National Biology Olympics for High School

SKILLS & SOFTWARE

- Programming languages: C++ (Std, Boost), Python, R (Bioconductor), Javascript (D3, jQuery), SQL, Perl, Ruby, Java

- Software: Visual Studio, Eclipse, Vim, Virtual Box, Hadoop, MySQL, Apache, Cytoscape, Git, VTK, FLUENT, Dreamweaver, Illustrator, Photoshop, AutoCAD

- Software project demonstration:
  - HS-SEQ: https://code.google.com/p/glycan-pipeline/
  - BRAIN: https://code.google.com/p/brain-isotopic-distribution/
  - MULTI-HS-SEQ: https://github.com/hh1985/multi_hs_seq
  - SpectrumAnnotation: https://github.com/lamarck2008/SpectraAnnotation