The Automation of Glycopeptide Discovery in High Throughput MS/MS Data

Sajani Swamy

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

Waterloo, Ontario, Canada, 2004

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

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

Abstract

Glycosylation, the addition of one or more carbohydrates molecules to a protein, is crucial for many cellular processes. Aberrant glycosylation is a key marker for various diseases such as cancer and rheumatoid arthritis. It has also recently been discovered that glycosylation is important in the ability of the Human Immunodeficiency Virus (HIV) to evade recognition by the immune system. Given the importance of glycosylation in disease, major efforts are underway in life science research to investigate the glycome, the entire glycosylation profile of an organelle, cell or tissue type. To date, little bioinformatics research has been performed in glycomics due to the complexity of glycan structures and the low throughput of carbohydrate analysis.

Recent advances in mass spectrometry (MS) have greatly facilitated the analysis of the glycome. Increasingly, this technology is preferred over traditional methods of carbohydrate analysis which are often laborious and unsuitable for low abundance glycoproteins. When subject to mass spectrometry with collision-induced dissociation, glycopeptides produce characteristic MS/MS spectra that can be detected by visual inspection. However, given the high volume of data output from proteome studies today, manually searching for glycopeptides is an impractical task. In this thesis, we present a tool to automate the identification of glycopeptide spectra from MS/MS data. Further, we discuss some methodologies to automate the elucidation of the structure of the carbohydrate moiety of glycopeptides by adapting traditional MS/MS ion searching techniques employed in peptide sequence determination. MS/MS ion searching, a common technique in proteomics, aims to interpret MS/MS spectra by correlating structures from a database to the patterns represented in the spectrum.

The tool was tested on high throughput proteomics data and was shown to identify 97% of all glycopeptides present in the test data. Further, the tool assigned correct carbohydrate structures to many of these glycopeptide MS/MS spectra. Applications of the tool in a proteomics environment for the analysis of glycopeptide expression in cancer tissue will also be presented.

Acknowledgements

I would like to acknowledgements several people without whom this project would have been possible. Thanks goes Caprion Pharmaceuticals and my supervisor Dr. Paul Kearney for funding this project. This project would also never have been possible without the scientific direction of Navdeep Jaitly, Dr. Alexandra Furtos-Matei, and Dr. Pierre Thibault. I would also like to thank John Tsang, Michael Hu and Jonathan Badger for insightful discussions about various aspects of bioinformatics.

I am deeply indebted to my family and friends for their endless support and for words of wisdom and encouragement. I would also like to thank Therese Biedl and Eowyn Cenek for helping me through some difficult course work. Finally, thanks also goes to Deepa Menon for help when I needed it the most, and to My Waterloo.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The advent of highly sensitive analytical techniques and computer technology has provided life science researchers with the ability to analyze the biology of an organism in its entirety. This type of large scale biological analysis has enabled research such as comparative expression analysis, where the entire gene or protein expression profiles of two individuals, for example a healthy patient and a patient carrying a specific disease, can be compared. It is hoped that such research will provide biologists with the information necessary to identify the key proteins involved in disease pathways, and ultimately, identify drug targets.

With the completion of several genome sequencing projects, the next major goal of life science research is to utilize genomic data to study the corresponding protein complement, or proteome. Proteomics today consists of two main goals: to identify all the proteins of a particular organism and secondly, to characterize these proteins in terms of structure (structural proteomics) and function (functional proteomics)[18]. The proteome however is more complex than the genome since it is a dynamic entity; it can change with the state of development, the tissue or even the environmental conditions of an organism. In addition, phenomena like alternate splicing[1] and post-translational modifications (section 1.2), can lead to a vast array of protein products, thus challenging the 'one-gene-one-protein' classical dogma of biology proposed by Beadle and Tatum[30]. The proteome is larger than the genome, and its analysis is more complex. This flow of information in life science research can be illustrated schematically as seen in figure 1.1.

1.1 Introduction to High Throughput Proteomics

Recent developments in technology have facilitated the analysis of hundreds of genes or proteins simulataneously (see figure 1.1). Improvements in both the sensitivity and the resolution of two main technologies, gel electrophoresis and mass spectrometry (MS), for the tasks of protein separation and identification respectively, have enabled the field of proteomics to quickly evolve. Only with these high throughput techniques, can we understand how multiple proteins are regulated together and over various time points. In addition, the increased sensitivity enables the analysis of proteins expressed in small quantities. These low abundance proteins, which are often biologically interesting, are difficult to analyze since proteins cannot be amplified like nucleic acids. In this section, we will describe these techniques, as well as how they are applied in proteome research.

[1] Post-processing of mRNA involves 'splicing' out various portions to create a final product. Often, this splicing step is performed at different locations along the gene, leading to a series of mRNA species.

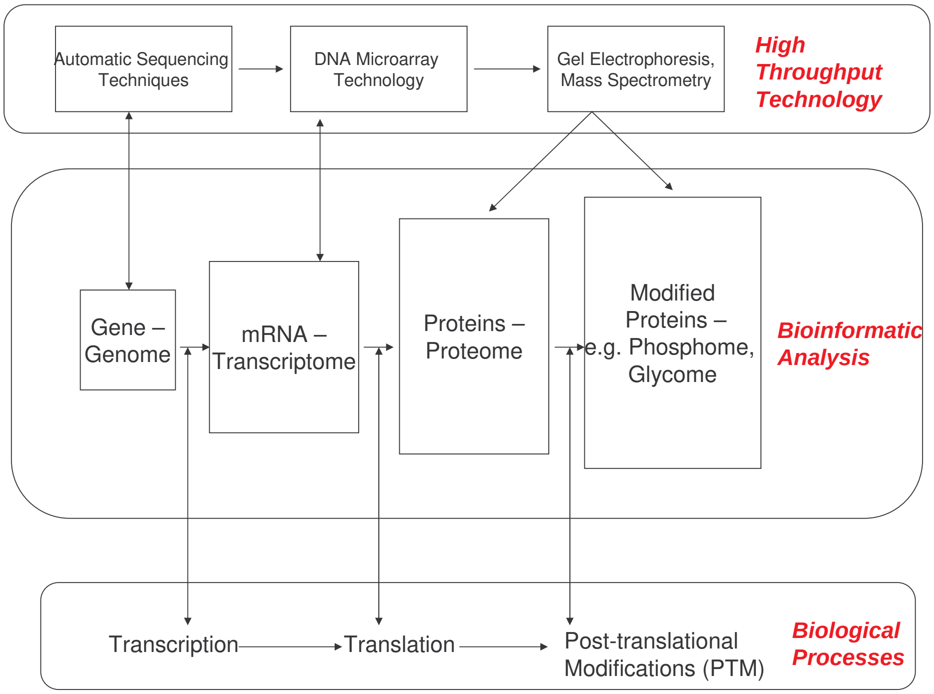

Automatic Sequencing Techniques
DNA Microarray Technology
Gel Electrophoresis, Mass Spectrometry
High Throughput Technology
Gene – Genome
mRNA – Transcriptome
Proteins – Proteome
Modified Proteins – e.g. Phosphome, Glycome
Bioinformatic Analysis
Transcription
Translation
Post-translational Modifications (PTM)
Biological Processes

Figure 1.1: Large scale life science research. The development of high throughput technology has enabled the analysis of the entire biological complement of an organism. At each level of the flow of biological information, there are several corresponding high throughput technologies to perform large scale analysis (high throughput technology flow), and a specific area of bioinformatics dedicated to its study (bioinformatics flow). The increase in size of the boxes in the bioinformatics flow pipeline (not to scale) demonstrate the increase in the complexity and volume of the data involved in its analysis.

1.1.1 Proteome Research Strategy

Today, there is a widely utilized strategy employed in high throughput proteome research (Fig.1.2). Typically, samples obtained from different cellular fractions are treated with an enzyme, usually trypsin, to cut them into smaller pieces or peptides, which are easier to analyze than full proteins. To further separate these peptides, they are processed by gel electrophoresis and the resulting images are electronically retrieved by high resolution scanners and analyzed using pattern recognition techniques to create proteome maps. The proteins in the gel are subsequently excised and treated, and may be further resolved using Liquid Chromatography (LC). To characterize the separated proteins, in terms of mass and peptide sequence, the samples are then usually subject to one or more rounds of mass spectrometry, typically Electrospray Ionization (ESI) or matrix-assisted laser desorption ionization (MALDI). The output from the mass spectrometer, MS/MS spectra, are subsequently analyzed with sophisticated bioinformatics tools and technology. This process can be visualized as shown in figure 1.2. In the next sections, these techniques will be further described.

1.1.2 Protein Separation

The first step in the proteome research strategy is peptide separation, which is achieved primarily by gel electrophoresis[2]. This technique allows the simultaneous separation of thousands of proteins according to various chemical properties. The results of the separation can then be viewed as a peptide 'map'.

Some types of electrophoresis such as two-dimensional electrophoresis,

[2] In this thesis, we will use gel electrophoresis as a general term to encompass various kinds of electrophoresis such as 1D, 2D, 2D-PAGE, 2D-SDS PAGE and so on.

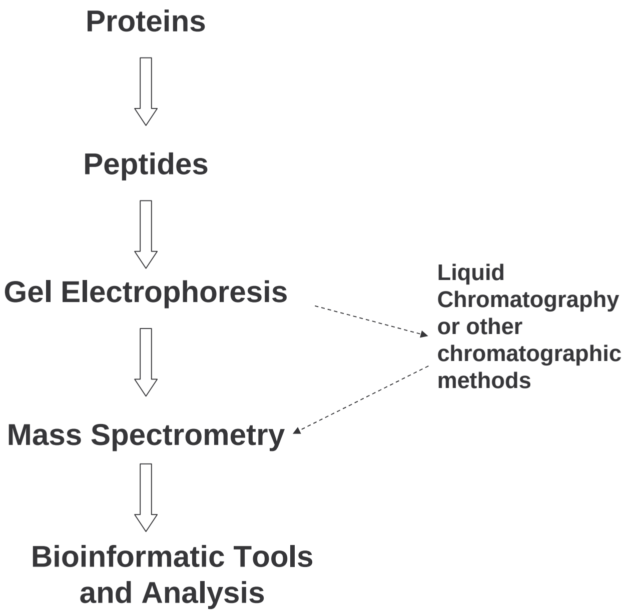

Proteins
Peptides
Gel Electrophoresis
Liquid Chromatography or other chromatographic methods
Mass Spectrometry
Bioinformatic Tools and Analysis

Figure 1.2: Typical strategy employed in proteome research (see text for details). Figure adapted from Bakthiar and Tse (2000).

can resolve more than 10000 proteins simultaneously in a highly reproducible way[14]. In addition, they can readily differentiate amongst many post-translationally modified forms of a protein[14]. The weaknesses of 2D-electrophoresis however is its inability to deal with certain classes of proteins, such as highly hydrophobic proteins, very small proteins, and those with isoelectric points at either extreme of the pH scale[14].

For protein separation, liquid chromatography (LC) is also commonly used. In liquid chromatography, proteins are dissolved in a liquid phase, and subsequently passed through several columns which separate the proteins on a number of dimensions. Separation on the basis of hydrophobicity is commonly employed in proteome research[2]. LC can be used for the direct analysis of the samples, but increasingly, is being used in tandem with 2D-electrophoresis to further concentrate proteins[29].

1.1.3 Protein Characterization - Mass Spectrometry

After separation, protein characterization is achieved by using mass spectrometry in most proteome projects (Fig.1.2). The goal of this technique is to provide information about the mass and chemical composition of a specific peptide. Developments of certain MS technologies such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) have made it possible to ionize and analyze large biomolecules[16]. Combined with improved apparatus design and refinements in sample preparation methods, mass spectrometry enables sensitive mass detection of the order of femtomoles[3]. Such sensitivity has made mass spectrometry a popular method for

[3] The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon-12; its symbol is ”mol.” A femtomole represents $10_{-}15$ of a mole.

analysis in a number of fields; it has been particularly successful in facilitating proteome research[16].

There are two main phases of mass spectrometry analysis of peptides. In the first phase, MS, the mass and abundance of each peptide entering the machine is measured. The output, a survey scan, lists all the peptides with unique mass-to-charge (m/z) ratios. When peptides undergo MS, they obtain one or more positive charges (z). The number of positive charges each peptide obtains is based on various parameters used on the operation of the mass spectrometer and the presence of various chemical groups on the amino acids of the polypeptide.

The most abundant peptides obtained from MS are subsequently selected for the second phase of mass spectrometry analysis, MS/MS[4]. Selected peptides are bombarded with enough energy from the ion source to cause its ionization and fragmentation into smaller peptides. Given that several identical peptides will enter the mass spectrometer for MS/MS when bombarded, an array of fragmentation products are produced each representing breakage at a specific bond[5]. The mass-to-charge ratios of the resulting fragments are then displayed as a spectrum in which peaks are drawn at various m/z points with heights proportional to the number of identical fragments observed at this m/z value (Fig.1.3). By examining the distances between the peaks of the MS/MS spectrum, the sequence of the peptide can be reconstructed. The resulting peak spectrum representing the composition of the molecule, can thus be used as a "molecular fingerprint". The process of MS/MS spectra interpretation for peptides is further discussed in chapter 2.1.

[4] Other criteria are also used to select peptides for MS/MS such as directed inclusion

[5] Not all peptide bonds are liable to break with the same probability.

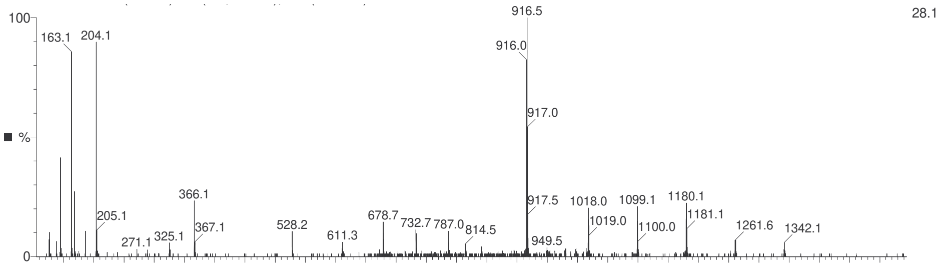


Figure 1.3: An example of an MS/MS spectrum. The labeled peaks in the spectrum represent the masses of peptide fragments. The x-axis represent the mass-to-charge (m/z) and the y-axis represent the relative abundance of that peak detected by the mass spectrometer. By calculating the distances between the peaks, the order of the peptide's fragmentation and thus the peptide's sequence can be inferred.

1.1.4 Bioinformatics - Database technology and analytical tools

After acquiring MS/MS spectra, these data are passed to various bioinformatics tools for data storage and analysis. In proteome research, databases play a crucial role. Some existing bioinformatic techniques for the analysis of proteomics data are further described in section 2.2.1.

1.2 Post-translational Modifications and Glycosylation

The previous section described details of proteomics research and how recent technology has enabled the simultaneous analysis of several hundred proteins to help elucidate complex disease pathways and protein interactions. In many cases however, protein analysis alone is insufficient to accomplish the goals of proteome studies. One factor that must be given special consideration is protein Post-Translational Modification or PTM. Protein PTM, the chemical modification of one or more amino acids in a protein chain, is a common phenomenon. There are a large number of known PTMs which occur in varying frequencies and have a wide range of roles in the alteration of protein structure and function. PTMs are crucial in the structure and function of many proteins and in the control of biochemical pathways[30]. Amongst the most ubiquitous and important PTMs is glycosylation.

Glycosylation As part of post-translation processing, proteins are often modified with the addition of one of more carbohydrate structures (sugars or glycans) forming glycoproteins[30]. The carbohydrate moiety of glycoproteins, called oligosaccharides or glycans, are composed of individual sugar residues or monosaccharides, which are covalently bonded together by glycosidic bonds. Recent estimates suggest that glycosylation affects over 60% of all proteins[26]. Protein glycosylation is crucial for many cellular processes including cell-cell interactions, protein-protein interactions, protein folding and trafficking, and is especially vital to the cell surface where it plays a key role in the proper positioning and functioning of surface receptors[30].

Glycosylation also plays a key role in disease pathways and examination of the glycome, or total glycoprotein complement of a cell, tissue or organism, is increasingly a priority in proteomics. Like the proteome, the glycome is dynamic, and reflects the physiological state of a cell. Glycosylation patterns can mirror biological processes taking place inside the cell and can alter with disease[26]. In rheumatoid arthritis for example, the levels of fully galactosylated[6] sugars decrease with disease activity[26]. In addition, aberrant glycosylation profiles are a key diagnostic marker of certain types of cancer. In a study of breast cancer glycosylation patterns, Whitehouse et al. [32] showed that breast cancer tumor cells often had MUC1 mucins [7] which carried shorter and less complex glycans compared to normal cells. Recently, viral coat glycosylation was also shown to be a key factor in the ability of the Human Immunodeficiency Virus (HIV), responsible for the Acquired Immunodeficiency Syndrome (AIDS) pandemic, to evade recognition by the immune system[31].

Glycoprotein Strucure The carbohydrate moieties of glycoproteins are attached to specific amino acid residues on peptides. There are two main

[6] Glycoproteins with attached galactose monosaccharide

[7] Mucins are large glycoproteins that carry many O-glycans

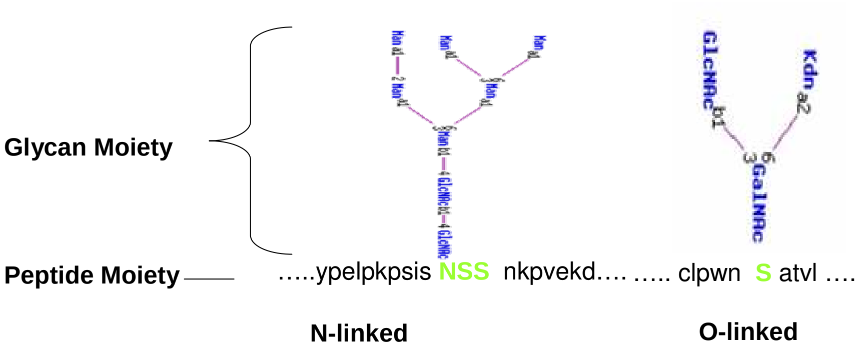


Figure 1.4: Glycoproteins are synthesized in two main varieties. N-linked glycoproteins are characterized mainly by carbohydrate attachment to an Asparagine amino acid (N), while carbohydrate attachment in O-linked glycoproteins is either to a Serine (S) or a Threonine (T) residue.

classes of glycoproteins, O-linked and N-linked, based on the site of attachment of the carbohydrate to the polypeptide chain. N-linked glycans are attached to the protein only at the amino acid sequence NXS/T[8] and are covalently linked to the N residue. O-linked glycans are linked to any serine or threonine amino acid in the polypeptide chain (Fig.1.4).

[8] N=asparagine, X=any amino acid, S=serine and T=threonine

Carbohydrate Structure The glycan moiety of glycopeptides are composed of monosaccharide residues. There are approximately 20 monosaccharide residues of which 6 are commonly seen (see table 4.5.1). Glycans consist of one or more monosaccharide residues bonded together by glycosidic bonds to form polysaccharide chains. The glycan moieties often have very complex structures and can be linear or highly branched depending on the pathways used in its synthesis.

Depending on the type of glycan, O-linked or N-linked, the structure of the glycans can vary greatly in terms of both size [9], structure and composition. The glycans found on mammalian N-linked glycoproteins have a common core composition of $\text{HexNAc}_2\text{Hex}_3$ [10] as illustrated in figure 1.5. Depending on the nature of the oligosaccharide chain attached to the common core, the glycan can further be classified as being oligomannose (or high-mannose), complex or hybrid. Oligomannose glycans (Fig.1.5A) contain only mannose residues, whereas complex type glycans have varied composition and a variable number of antannae stemming from the core (Fig.1.5B) [10]. Complex type N-glycans show the largest structural variation resulting from the combination of monosaccharides and the different number of antannae. Hybrid-type N-glycans have the characteristic features of both complex-type and high-mannose type glycans as seen in figure 1.5C [10]. The structures of O-linked glycans is less defined than that of the N-linked glycans. O-linked glycans are composed of six different cores, and can have varied structures and compositions stemming from the cores[10].

[9] Sizes can range from di-, and tri-saccharides to hundreds of saccharides.

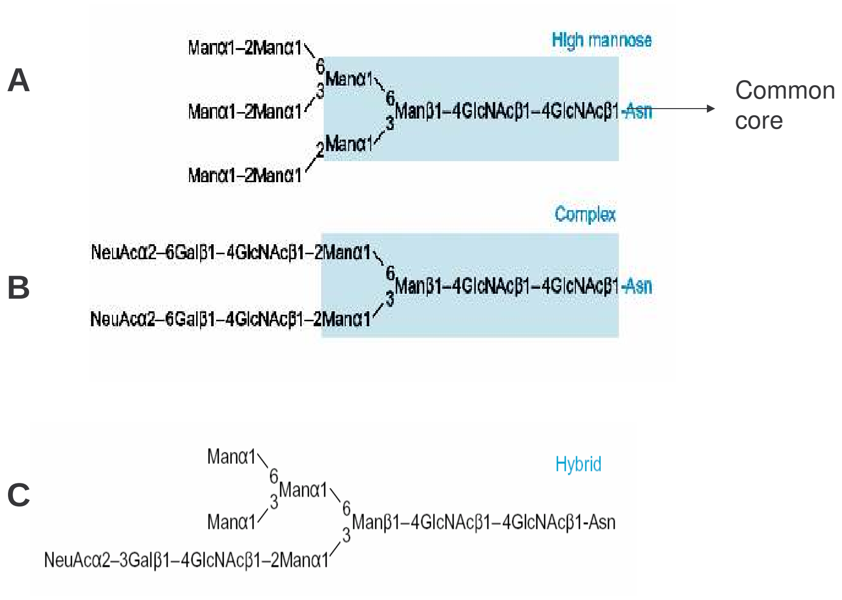

A
High mannose
Manα1–2Manα1
6
Manα1
3
Manα1–2Manα1
6
Manβ1–4GlcNAcβ1–4GlcNAcβ1-Asn
3
2
Manα1
Manα1–2Manα1
Common core
Complex
B
NeuAcα2–6Galβ1–4GlcNAcβ1–2Manα1
6
Manβ1–4GlcNAcβ1–4GlcNAcβ1-Asn
3
NeuAcα2–6Galβ1–4GlcNAcβ1–2Manα1
C
Hybrid
Manα1
6
Manα1
3
Manα1
6
Manβ1–4GlcNAcβ1–4GlcNAcβ1-Asn
3
NeuAcα2–3Galβ1–4GlcNAcβ1–2Manα1

Figure 1.5: The glycans of mammalian N-linked glycoproteins contain a common core structure of HexNAc$_2$Hex$_3$ as shown in this figure. Depending on the structure stemming from the common core structure, a glycan can be classified as oligomannose (A), complex (B) or complex (C).

1.3 Introduction to Glycomics

Despite the importance of glycosylation in our understanding of diseases, the analysis of protein glycosylation remains a low-throughput and laborious task. One reason that protein glycosylation is difficult to analyze is that protein glycosylation is a complex process and glycoproteins have highly heterogeneous structures. Unlike the genetic code, there is no rigid template that accurately specifies glycosylation patterns, but rather a complex assembly-line system involving competition by hundreds of gene products[30]. It is not uncommon for a glycoprotein to be processed with more than 100 alternative glycans at a single glycosylation site[26].

The heterogeneity of glycoproteins arises at two main junctures in the glycoprotein structure (Fig.1.6):

- Glycosylation sites on the protein (macroheterogeneity). A protein can have one or more glycosylation sites, which may or may not be occupied by a carbohydrate.
- Monosaccharide composition of the glycan (microheterogeneity). A single glycosylation site can have a series of different carbohydrates, or glycoforms, attached

The complexity of glycoprotein analysis is further complicated by the fact that each cell, tissue, organ and organism exhibits different glycosylation patterns, which can change based on the cell's state or activity. The analysis of such immense variation in glycoprotein structure is very hard to substantiate, and multiple analytical methods are needed to fully characterize glycoproteins.

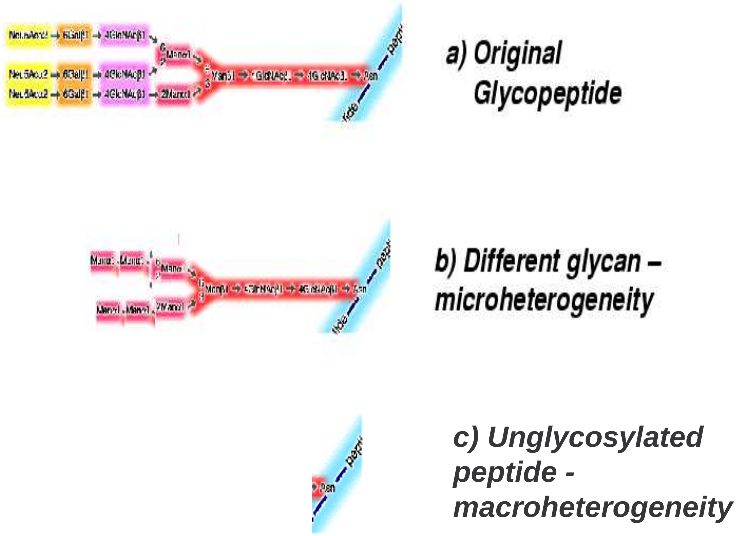

NeuAcα2
6Galβ1
4GlcNAcβ1
Manα1
2Manα1
Manβ1
GlcNAcβ1
Asn
peptide
a) Original Glycopeptide
b) Different glycan – microheterogeneity
c) Unglycosylated peptide - macroheterogeneity

Figure 1.6: Glycopeptide heterogeneity. There are two types of glycopeptide heterogeneity. This schematic illustrates a specific glycopeptide in 3a. The coloured boxes represent the various monosaccharides comprising the glycan. b)Example of microheterogeneity; this glycopeptide contains only mannose sugars unlike 3a. The glycans in 3a and 3b are therefore glycoforms. c) Example of macroheterogeneity; the glycosylation site of the original glycopeptide is not occupied.

In addition to structural complexity, glycoproteins are difficult to analyze from an experimental standpoint. To fully capture and elucidate the glycan structure of a glycoprotein involves several analytical steps that often require a large amount of sample and which can be costly. Problems in glycoprotein analysis are further described in section 1.3.5.

1.3.1 Glycome Research

Today many of the high throughput technologies used in proteomics are also applicable to glycomics, and there have been several large scale glycomics initiatives put forth. Researchers at Teikyo University in Japan for example, have begun a project aimed at elucidating the glycome of *Caenorhabditis elegans*, whose genome has already been completely sequenced and annotated[17]. Similar to high throughput proteomics, glycomic projects consist of two main steps : i) glycoprotein isolation, and ii) glycoprotein identification. Although similar concepts are applied, the diversity and complexity of glycoproteins require different methodologies for analysis (Fig.1.7).

1.3.2 Glycopeptide Isolation

In terms of separation techniques, electrophoresis is a commonly used technique when studying glycopeptides (Fig.1.7). Electrophoresis permits the separation of different glycoforms, which are typically represented by one or more diffuse bands during gel electrophoresis[30]. To isolate and visualize the glycoproteins from the gel, there are several staining reagents that can be used, the most common ones being lectin staining and immuno-staining. Lectins are specialized glycoproteins for carbohydrate-binding[30]. They are commonly used as probes to bind and isolate carbohydrate-containing molecules

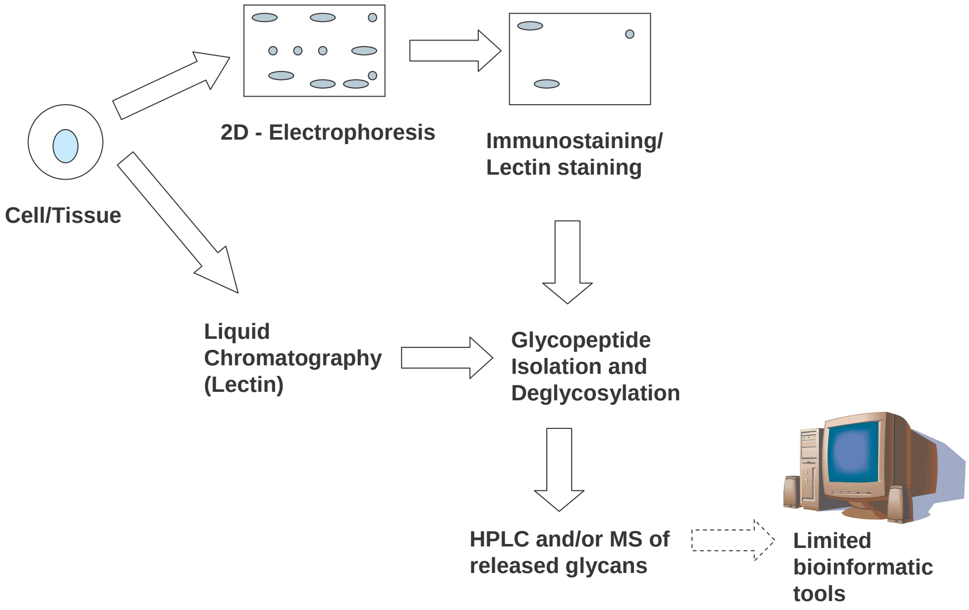


Figure 1.7: Common strategy employed in glycome research (see text for details).

and when applied to gels, will bind and stain the gel where glycoproteins exist[30]. Alternately, carbohydrate-recognizing antibodies can be used to bind and visualize the glycoproteins, a process called immunoblotting staining.

In many situations, LC methods are preferred to electrophoresis. For example, some glycoconjugates do not enter ordinary gels or migrate only as smears[30]. In these situations, LC, specifically lectin-affinity chromatography is often preferred to 2D-electrophoresis. Chromatography columns can be specially prepared with lectin to permit the binding of the carbohydrates and thus glycosylated proteins. Furthermore, the lectin columns can be designed to isolate a particular type of glycan for added specificity [30].

Once the glycoproteins are identified and isolated, they are deglycosylated using chemical cleaving agents. The peptide moieties can further be identified using protein characterization techniques discussed in the previous section, and the released glycans characterized as described below.

1.3.3 Glycopeptide Characterization

The characterization of the released glycans entails determining three main features: i) monosaccharide composition, ii) protein attachment site and iii) sequence branching and linkage positions [10]. Like in proteomics, advances in mass spectrometry have made this technology a popular choice for glycopeptide analysis. When soft-ionization techniques are applied to treated glycoproteins, MS analysis is capable of analyzing mixtures of oligosaccharides and provides information about oligosaccharide molecular weight, the heterogeneity of a sample and structural information such as branching[30].

[10] Determining the anomeric configurations of the monosaccharides is often another characteristic examined in glycopeptide characterization

Although not matching the level of sensitivity currently enjoyed by peptide analysis (femtomole), advanced MS instrumentation is now capable of mass profiling glycans at relatively high resolution and sensitivity (picomole), giving routine mass accuracy of 50-100 ppm[11][30].

1.3.4 Bioinformatics Tools and Databases for Glycome Research

Glycomics is a new and emerging field, and as such, there are very few publicly available glycomics databases. GlycoSuiteDB[12] is among the few databases available publicly on the Internet since early 2001. It is a relational database that curates information from scientific literature on glycoprotein derived glycan structures[6]. Since the structures registered in GlycoSuiteDB are only from reported glycans, there are a very small number of entries so far. Another publicly available glycome database is O-Glycbase curated by the Centre for Biological Sequencing from the Technical University of Denmark. O-GLYCBASE is a database of only O-glycosylated proteins[19] which have at least one experimentally verified O-glycosylation site. O-GLYCBASE is coordinated with NetOGlyc, a tool for predicting O-linked glycosylation sites[19]. Both GlycoSuiteDB and O-GLYCBASE are extensively cross-linked to various nucleotide and protein databases.

In terms of glycomic tools, there are also a relatively limited number of publicly available tools. One tool, GlycoMod, is a tool that predicts possible oligosaccharide structures from their experimentally determined masses[5]. Several other algorithms for automated glycan MS/MS spectra interpretation are discussed in section 4.2. As glycomics matures as a field, there will

[11] A measure of concentration. 1 ppm = 1 mg/liter. One ppm is 1 part in 1x10^6.

[12] http://www.glycosuitedb.com

likely be more databases and analysis tools available as currently exist in proteomics.

1.3.5 Technical and Industrial Problems in Glycomics

Problems in Traditional Glycoprotein Analysis Glycoprotein analysis is still difficult despite the development of several analytical techniques. In terms of industrial processing, glycoprotein analysis is a time-consuming, costly process and low-throughput method due to the due complexity and diversity of glycan structures. Glycoprotein analysis is also unsuitable in some situations as traditional methods of analysis involve chemical deglycosylation. Deglycosylation requires a large sample amount to be effective which may not be available in research projects where tissue samples are limited. Moreover even without sample limitations, deglycosylation is often unfavourable as the treatments usually damage the peptide precursors precluding further analysis of the peptide moiety for glycoprotein identification [30].

The development of several soft-ionization techniques has provided technology that can circumvent the problems mentioned above. Fast Atom Bombardment (FAB-MS), ESI-MS and MALDI-MS permit the analysis of intact glycoconjugates without requiring chemical treatment[8].

Problems in Glycoprotein Analysis in Proteomic Data It is often desirable to concomitantly analyze proteins and glycoproteins in a single proteomic project. As mentioned above, samples for proteomic analysis are often limited and there may not be enough sample for analysis geared towards unmodified proteins followed by analysis specifically for glycoproteins. Since ionization techniques used in proteomics are also applied in glycomics, it is possible to analyze proteins and glycoproteins simultaneously. However, this concomitant analysis may not always be effective due to different chemical characteristics and mass differences between glycoproteins and proteins. When the parameters for MS analysis have been geared towards protein analysis, the carbohydrate groups of the glycopeptides mitigates against their selection for MS/MS. For example, the upper mass limits of the mass spectrometer are often too low for glycopeptides, which generally have a greater mass than peptides. In addition, LC techniques employed for protein analysis may not be appropriate for glycoprotein analysis which generally require a longer elution time given their hydrophilicity. Despite these problems, the analysis of captured glycopeptides in proteomic data is important and can still provide valuable biological information.

1.4 Research Problem

In proteome studies today, the analysis of glycopeptides is an increasingly crucial task as researchers discover the link between aberrant glycosylation profiles and disease. Since many of the glycopeptides of interest are often found in low abundance, sensitive methods with low sample requirements are necessary for analysis. In light of this requirement, the preferred technique for glycopeptide analysis is mass spectrometry as described in Section 1.3. Given the high volume of data output from proteome studies however, it is impractical to manually detect and characterize all glycopeptides present. As a result, there is a pressing need to automate glycopeptide discovery. In this thesis we address the following problem:

Is it possible to detect, identify and characterize glycopeptides from high throughput mass spectral data in a rapid and accurate manner?

In our attempt to solve this problem, we will consider two sub-problems:

- How to identify glycopeptides from MS/MS spectra (Glycopeptide Classification)
- How to elucidate the structure of the carbohydrate moiety of the identified glycopeptides (Glycan Analysis)

Each of these sub-problems will be presented in more detail in the following chapters of this thesis.

1.5 Thesis Overview

In this thesis, we propose a methodology for the automatic identification and characterization of glycoproteins from ESI-MS/MS data. In Chapter 2, we discuss some of the problems involved in automatic MS/MS spectra interpretation, and present some existing software and computational approaches available for this task in peptide analysis. In Chapter 3, we present a technique for the identification of glycopeptide MS/MS spectra in proteomic data. Chapter 4 presents an automated method for the analysis and characterization of the glycan moiety of glycopeptides. Finally, we present our results and conclusions in chapters 5 and 6 respectively.

Chapter 2

Automated Analysis of MS/MS Data in Proteomics

In this section, we will examine the basics of MS/MS spectra interpretation and review some existing algorithms and techniques for automated MS/MS spectra interpretation.

2.1 Interpretation of Peptide MS/MS Spectra

The process of MS/MS spectra interpretation is a time-consuming one and requires the experience of a trained expert. Spectral interpretation for peptides and proteins entails the analysis of the peaks of the spectrum and the identification of the molecular fragments they represent. The entire molecule can be reconstructed by observing the order in which the various fragments appear.

Peptides are made of repeating amino acid units as illustrated in figure 2.1. To differentiate the start of the peptide from the end, each terminus of the peptide is labeled; one terminus is called the N-terminus (amino terminus) and the other the C-terminus (carboxyl terminus)[1]. When the amino acids of the peptide fragment, they produce 2 types of ions: a b-type and a y-type ion[2]. A b-ion and a y-ion represent the peptide fragment containing the N- or C-terminus respectively[3]. Every b- and y-ion pair is complementary and the combined masses of the b-ion and the y-ion should equal the mass of the peptide. The interpretation of peptide MS/MS spectra thus begins with the identification of a series of ions (b or y), between which there is a mass difference equal to a specific amino acid. By detecting the entire series of b- or y-ions and observing mass differences between them, the sequence of the peptide can be reconstructed[28]. This process is illustrated in figure 2.2.

The process of reconstructing peptide sequence by observing peak differences in the spectrum, *de novo* sequencing, is difficult since peptide spectra are complex. The following section describes some of the sources of complexity in spectra interpretation, as well as how to recognize and manage them.

2.1.1 Confounding Factors in Spectra Interpretation

Even in the absence of experimental error, MS/MS spectra can contain a great deal of ambiguity as a result of complex molecular fragmentation patterns. To fully and correctly reconstruct molecular structures from MS/MS spectra requires careful consideration of several factors which are further

[1] The N terminus is where protein synthesis is initiated and the C terminus is where it is terminated.

[2] There exist other types of ions as well such as x-ions and a-ions although they are more rare.

[3] The appearance of a b-type ion or y-type ion depends on which of the fragments obtained a positive charge during the fragmentation and ionization processes.

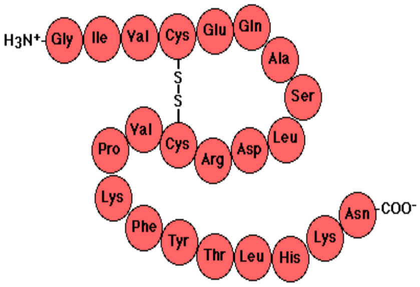


Figure 2.1: This figure illustrates the structure of a protein. Proteins are made up of repeating units of amino acids. There are 20 common amino acids found in nature. In this figure, each circle represents one amino acid unit which together form a polypeptide chain. In this diagram, interactions between the various amino acids are also illustrated. These interactions are partially responsible for conferring a three-dimensional structure to proteins.

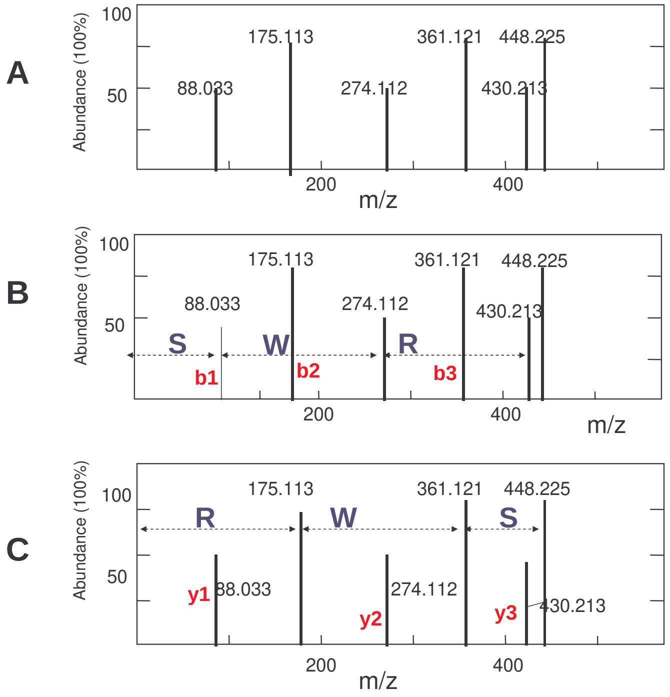


Figure 2.2: This figure illustrates the process of determining peptide sequence from MS/MS spectra. Part A shows an MS/MS spectrum for the sequence SWR. To determine the peptide sequence requires the identification of either the b-type or y-type peaks series shown in figures B and C respectively. Once these series are identified, the peptide sequence can be elucidated by examining the distances between the peaks and correlating these distances to known amino acid masses.

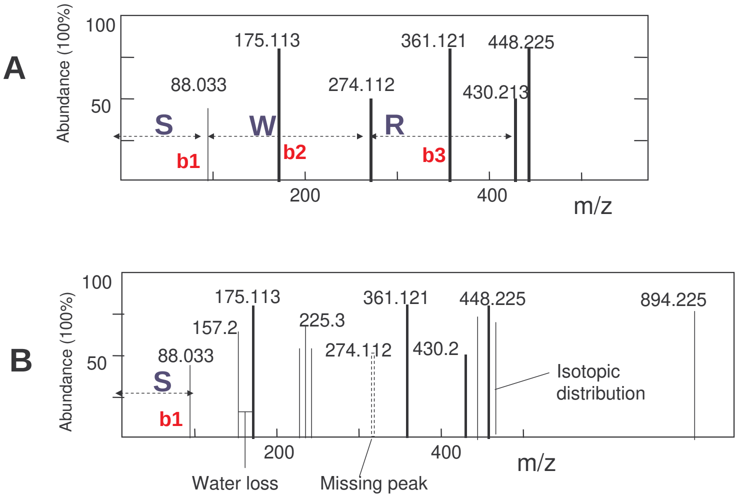


Figure 2.3: This figure illustrates some of the complicating factors in the interpretation of peptide MS/MS spectra. Figure A represents an ideal spectrum for the peptide sequence SWR. In B, some complicating features of MS/MS spectra such as water loss, missing peaks and multiply charged peaks are illustrated (see text below for additional explanations).

described below. Figure 2.3 illustrates a realistic peptide spectrum with examples of some of these complicating factors.

Missing Peaks It is a common occurrence in MS/MS spectra that expected fragmentation peaks are missing. Although this phenomenon can be caused by experimental parameters used, it may also be the result of the inherent nature of a particular molecule under certain ionization conditions. For example, the fragmentation on the C-terminal side of proline is often absent or is of reduced intensity, and often there is often a lack of fragmentation between the first and second amino acids of peptides when using low-energy ionization[15]. In addition, in the case of peptides, certain amino acids can bias fragmentation reducing the abundance of some fragments significantly[15]. In figure 2.3b, the missing peak at m/z 274.112 causes the b-type ion series to be discontinuous complicating peptide sequence determination.

Isotopes Given the accuracy of mass spectrometers[4], the chemical isotopes of the carbon atoms of peptides may often be detected and a peak displayed for each isotope[5]. Singly charged peptide isotopes can be recognized by a specific pattern of adjacent peaks spaced by 1 mass unit[27]. Once isotopic peaks are identified, they can either be combined to represent an average of the peak values, or one specific isotope can be selected to represent the most common isotope.

Multiply Charged Species In the ionization of biomolecules, it is often the case that the ions produced exist in several charged states, and for which multiple peaks will be represented. It is a phenomenon that is particularly noted in ESI-MS spectra in which peptides often obtain +2 or +3 charges. In figure 2.3b, peak 448.225 is in reality a doubly charged peak whose real mass is at 894.225. These multiply charged peaks can be identified by the spac-

[4] Not all mass spectrometers provide this level of accuracy. For those commonly employed in proteomics such as Q-TOF machines, this level of accuracy is achieved.

[5] Isotopes are atoms of the same element with a different mass. Isotopes of a particular atom often have the same chemical attributes, but often display different physical attributes; e.g., carbon-12, which is stable, and carbon-14, which is radioactive. (http://www.academicpress.com/inscight/10211997/isotope2.htm)

ing between adjacent peaks[6]: doubly charged peaks have spaces of 0.5 and triply charged peaks have 0.333 daltons separating them. Removing multiply charged ions by converting them to a singly charged m/z, or *deconvoluting* the spectra, can greatly improve the overall clarity of the spectrum, making interpretation less complicated. There are several programs available for automatic spectral deconvolution which employ a variety of computational approaches[11].

Adducts Formed During Ionization Another factor that often confounds mass spectra interpretation is the presence of additional peaks from ions derived from side reactions with the sample fragments[15]. Neutral loss of water (Fig.2.3b), sodium or ammonia adducts, frequently occur during the ionization of certain peptide fragments[15]. For each fragment that undergoes such reactions, there are adjacent peaks in the MS/MS spectrum representing these adducts. In addition, spectra are frequently complicated by ions that are fragmentation products of fragmentation products[15]. If these doubly-ionized fragments are of a high enough intensity, they can be misleading during sequence determination, as they cannot be explained by any of the other fragmentation adducts, singly or multiply charged species.

Single units that are isobaric with coordinated units Another confounding factor in the interpretation of MS/MS spectra is that combinations of some of the lower mass molecules have identical or nearly identical masses of higher molecular weight molecules[15]. In the interpretation of carbohydrate spectra for example, one hexose residue and one NeuAc[7] residue

[6] These peaks represent the isotopes of the species. At charge 1, the isotopes are separated by 1 mass unit. Consequently, a doubly charged isotope will be separated by 0.5.

[7] NeuAc = N-acetyl neuraminic acid (sialic acid), NeuGc = N-glycolyl neuraminic acid

combine together to give the same mass (453.1482 Da) as one deoxyhexose residue plus one NeuGc13 residue (453.1482 Da)[5]. When interpreting a spectrum, it is possible to overlook a cleavage ion between such disaccharides and to incorrectly infer the presence of the corresponding higher mass monosaccharide. Conversely, a high mass monosaccharide could be construed as a disaccharide. Thus, all isobaric masses should be considered in spectra interpretation, adding to the complexity of the interpretation process.

2.2 Algorithms for Automated MS Interpretation

Given the high volume of data produced and the time involved in manual spectra interpretation, there have been several approaches put forth for automatic MS/MS spectra interpretation. In general, there are two main approaches to automated MS/MS interpretation. The first one involves the development of algorithms to simulate the process of manual spectral interpretation and analyzes the information contained in the spectra directly to derive putative peptide sequences. The other main strategy makes use of curated protein and nucleotide databases, and matches MS/MS spectra to database protein sequences.

2.2.1 Database-Searching Programs for Peptide Sequencing - MS/MS Ion Search

Since the idea of using databases for protein identification was first put forth in 1993, there have been many approaches developed for this task. MS/MS ion searching usually consists of an initial step, Peptide Mass Fingerprinting (PMF) which aims to identify the parent protein of peptides based on their experimentally derived mass. There are several approaches for this method including probabilistic techniques (Mascot) and techniques which score matches based on the length of the peptide such as MS-Fit. To further elucidate the amino acid sequence of the peptide, MS/MS ion searching techniques are applied[25].

There have been several approaches put forth for MS/MS ion searching. In general, they proceeds as follows (Fig.2.4):

- The database proteins are digested[8] *in silico*, and the theoretical masses of the peptides are obtained

- The experimental masses of the peptides obtained by MS are matched to their parent proteins using PMF Techniques.

- The database peptides which have been matched to experimental peptide are further fragmented *in silico* to produce a set of peaks corresponding to theoretical peptide ionization fragments.

- The observed spectra obtained by MS/MS are correlated to the virtual spectra by a variety of computational approaches, and a score assigned based on the quality of the match.

- Based on the score, if the observed peptide is matched to a sequence in the database, the corresponding entry is extracted. If the sequence database does not contain the observed protein, many programs identify those entries which exhibit the closest homology, often equivalent proteins from related species.

Most of the MS/MS ion searches differ in the scoring scheme for the correlation of the virtual and experimental spectra. SEQUEST developed

[8] Usually with trypsin.

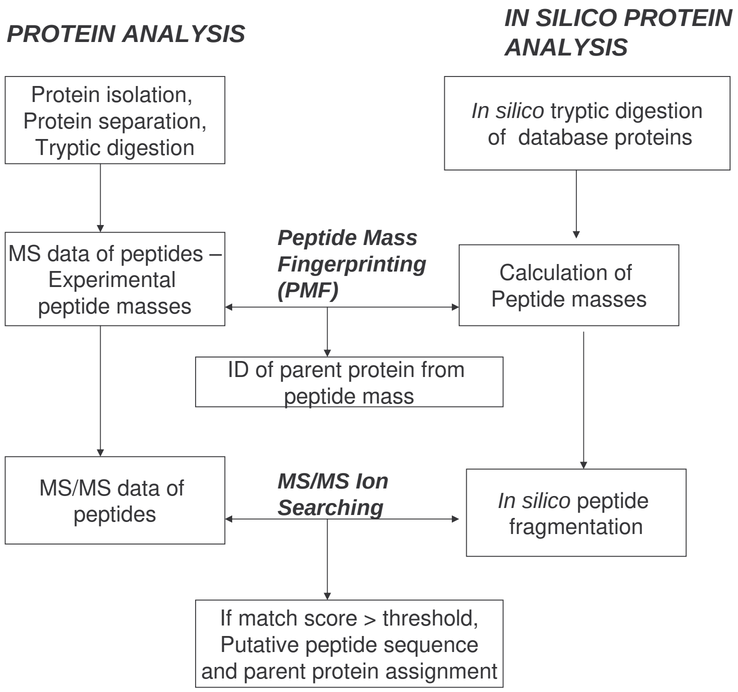

PROTEIN ANALYSIS
IN SILICO PROTEIN ANALYSIS
Protein isolation, Protein separation, Tryptic digestion
In silico tryptic digestion of database proteins
MS data of peptides – Experimental peptide masses
Peptide Mass Fingerprinting (PMF)
Calculation of Peptide masses
ID of parent protein from peptide mass
MS/MS data of peptides
MS/MS Ion Searching
In silico peptide fragmentation
If match score > threshold, Putative peptide sequence and parent protein assignment

Figure 2.4: Database searching programs for peptide sequencing. The usage of database technology for protein identification and characterization consists of two main methods : Peptide Mass Fingerprinting (PMF) and MS/MS ion searching. This schematic illustrates the general approach used in these methods of which there are many variants. See text for details.

by Yates et al, matches observed spectra to virtual spectra by using Fast Fourier transforms [7]. Another program developed by Matrix Science, MASCOT, uses a probability-based scoring scheme to evaluate the quality of a match. The fundamental approach is to calculate the probability that the observed match between the experimental data set and each sequence database entry is a chance event[25]. The match with the lowest probability is reported as the best match. There are several other programs such as MS-Tag, PepFrag and PeptideSearch, all which use different approaches to address this problem[23].

Evaluation of method The benefits of database searching techniques are numerous. They provide accurate peptide sequence elucidation and enable the linking of experimental data to database information. In addition, this method is fairly robust against missing peaks and other obfuscating factors in MS/MS spectra analysis discussed in the previous section. This is mainly because the matches are determined based on the overall fitting of the peaks. PMF and MS/MS ion searches will likely improve with increases in the sensitivity of MS technology, and fewer peptides will be needed to for protein identification[7].

Despite the advantages of using database techniques for spectra interpretation, there are several shortcomings in PMF and MS/MS ion search techniques. For one, database searching cannot provide sequencing for proteins that are not included in the particular database, such as novel proteins or modified proteins[9]. This problem is especially relevant in proteomics as many novel proteins are discovered. In addition, *in silico* digestion may

[9] Since many protein databases are created by from genomic information, modifications are often not annotated.

not always be accurate for several reasons such as missed or non-specific cleavages, leading to false negatives[15]. Database techniques have also been criticized for not scaling up very well with the length of a protein and size of protein databases which are growing rapidly, since the probability of a random match increases[4]. Pruning techniques have been applied to avoid this situation, but it is at the cost of reduced accuracy[4]. Finally, protein databases are often incorrectly annotated and as such can report false results. A summary of these points is presented in Table 2.2.1.

Advantages	Disadvantages
• Permits protein identification and peptide sequencing • Provides link with biological databases and access to information contained in cross-linked databases • Somewhat robust against missing fragment peaks and other factors leading to MS/MS spectral complexity. • Technology will likely improve with increases in MS mass accuracy	• Cannot be used for novel proteins • As database size increases, may obtain more random matches. If MS accuracy is not proportionally increased, there could be a drop in the confidence level of the results • Theoretical digestion and fragmentation may not adequately predict peptide products (due to missed cleavages, post-translational modifications, or other reasons), missing potential matches. • Dependent on database accuracy, many proteomic databases contain errors • In general, cannot account for modified peptides

Table 2.2.1 Evaluation of MS/MS Ion Searching Techniques.

2.2.2 Algorithms for *de novo* sequencing

Apart from database techniques, there have been several algorithms developed to infer peptide sequences directly from the spectra themselves, simulating the process of manual *de novo* sequencing. There are several programs available that employ a variety of approaches such as dynamic programming (PEAKS[21]), and graph theoretic approaches (Papayannopoulous[24]).

Evaluation of Method De novo sequencing techniques provide solutions to many of the problems associated with database sequencing methods mentioned in the previous section. However, few of the *de novo* techniques have enjoyed widespread use or success. This is partially because de novo methods are highly dependent on the quality of the spectrum. Without much spectra pre-processing, which can be computationally expensive, these methods are generally not robust to noise and missing peaks. In addition, they do not provide protein identification directly, and must be coordinated with database techniques to determine the identity of the parent protein. A summary of the relative advantages and disadvantages of the *de novo* techniques is listed below in table 2.2.2

Advantages	Disadvantages
• Can provide information for novel proteins • Does not depend on database technology - scalable with increases in database size and accuracy	• Relies on good quality spectra; requires extensive spectra deconvolution for efficient sequencing

Table 2.2.1 Evaluation of automated *de novo* sequencing techniques

2.3 Automated Spectra Interpretation - Summary

In general, automated spectra interpretation is a difficult problem. Due to various obfuscating factors present in MS/MS spectra, it is often impossible to determine a peptide sequence with absolute confidence. However, even partial sequences derived automatically can provide useful sequence information.

The two techniques discussed above are appropriate in different situation, and each have specific strengths and weaknesses. In general, database searching techniques facilitate protein identification and are accurate when the observed peptide is found in the protein databases. However, when dealing with novel proteins *de novo* algorithms may be the preferred method. In addition, they do not rely on the accuracy of the searched databases.

A coordination of the two techniques is commonly used, and offer a method of automatic results validation. Since the two methods use quite different approaches, an agreement between them indicates a high likelihood of a putative sequence being correct.

Chapter 3

Glycopeptide Classification

The previous chapter provided an overview of automated mass spectra interpretation in proteomics. We discussed some of the challenges of mass spectra interpretation as well as a description of existing methods for automated analysis. In this chapter, we present our design for a tool for glycopeptide identification from high throughput ESI-MS/MS proteomic data.

3.1 Introduction

In problems involving classification, the main goal is to learn a mapping from a vector of measurements $\mathtt{x}$ to a categorical variable Y [13]. The categorical variable to be predicted will take values from the set C, c*1*, .., c*m*, where each c*i* represents a unique class of objects[13]. Thus, the role of the classifier is to take a input vector of attributes $\mathtt{x} = \{X1,.., Xp\}$, for p attributes and map these values to a score S which can be used to place $\mathtt{x}$ into one of m bins of C.

In the identification of glycosylated spectra from ESI-MS/MS data, we would like to classify spectra as being glycosylated or non-glycosylated. Formally the problem can be stated as follows:

Problem: Given an input spectrum, E*, with m attributes x = X1..Xm, map x to a score S. Subsequently, use S to classify E as a glycopeptide or non-glycopeptide.*

A solution to this problem involves:

- Defining a model for glycopeptide spectra. This step entails listing the attributes of glycopeptide spectra, and deriving a function for evaluating each attribute in the experimental spectrum E
- Defining a score S based on the results of each attribute function
- Defining a mapping from score S to one of two classes: glycopeptide or non-glycopeptide

These steps will be further discussed in this chapter.

3.2 Model for Glycopeptide ESI-MS/MS Spectra

In this section, we will discuss the features of glycosylated ESI-MS/MS spectra. From this analysis, it will be possible to define an attribute list for the glycopeptide model.

3.2.1 Glycopeptide Fragmentation

Glycopeptides fragment in a characteristic manner when subject to Collision-Induced Dissociation (CID) and display a recognizable signature in their

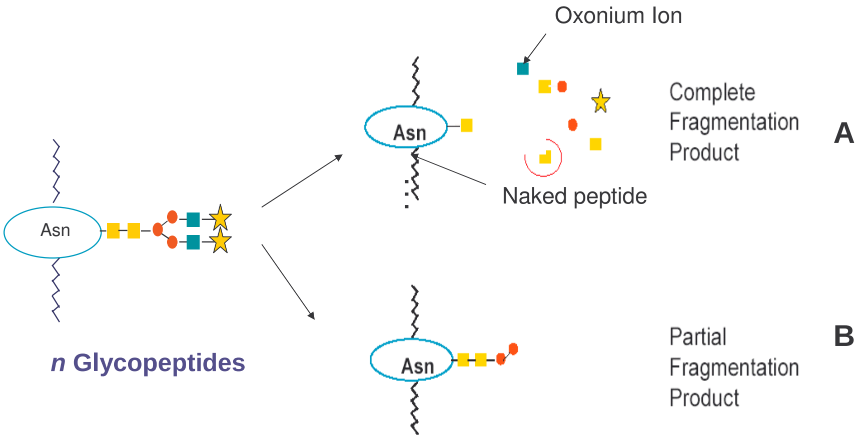


Figure 3.1: Schematic of N-linked glycopeptide fragmentation. Upon CID, the more labile carbohydrate appendage dissociates typically leaving a backbone peptide with the first GlcNAc residue still attached. Full glycan fragmentation as shown in A. In B a partial glycopeptide fragment is shown.

MS/MS spectra. The more labile glycosidic bonds of the carbohydrate moiety are broken and the peptide backbone remains unfragmented (Fig. 3.1a). The only monosaccharide of the glycan which does not usually fragment is the first GlcNAc residue linked to the peptide moiety since the β-glycosylamine linkage of GlcNAc to Asn is stronger than that of the glycosidic bonds (Fig. 3.1a). Upon the breakage of each glycosidic bond, two fragmentation products are produced: a low mass oxonium ion and a partially fragmented glycopeptide (Fig. 3.1b). Figure 3.1 illustrates the process of glycosidic bond breakage and complete glycopeptide fragmentation.

Since several copies of the same glycopeptide enter the MS/MS chamber simultaneously, after CID there will exist several species of glycopeptides, forming a mixture of various partially fragmented glycopeptides (Fig. 3.1b). Each partially fragmented glycopeptide, which has a unique mass-to-charge (mz) ratio, is registered by the mass spectrometer and a spectrum is produced illustrating the relative number of each type of fragment species.

3.2.2 Attributes of Glycopeptide Spectra

The unique fragmentation pattern of glycopeptides creates spectra which are identifiable by visual inspection. In figure 3.2, the general appearance of glycopeptide spectra is contrasted with those of random spectra (neither glycopeptide nor peptide) (Fig.3.2b), and peptide spectra (Fig.3.2c). In general, glycopeptide spectra contain three characteristic features which differentiate them from other types of MS/MS spectra: oxonium ions peaks, differential peak densities, and peaks spaced by various saccharide combinations in the high m/z range of the spectrum. A typical glycopeptide spectrum is illustrated in figure 3.3. Each of these features of glycopeptide spectra are further

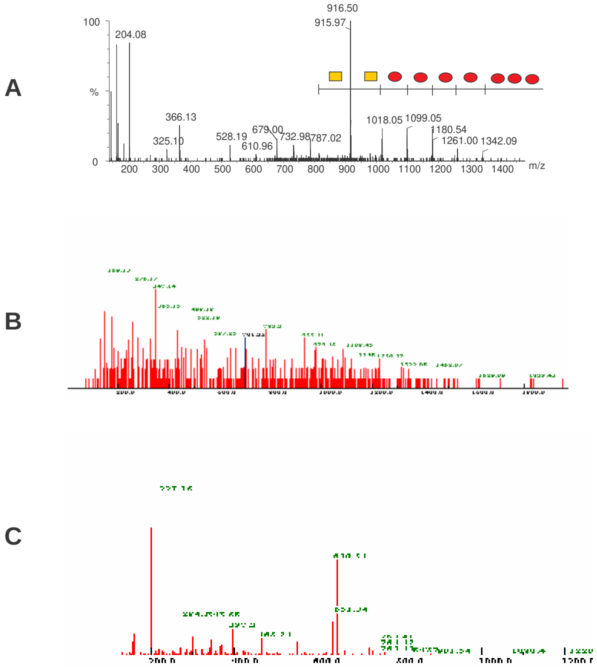


Figure 3.2: This figure illustrates the differences in the general appearance of the spectra, in terms of peak distributions and intensities, between glycopeptides (A), random (neither glycopeptide nor peptide) (B) and peptides (C).

described below.

Oxonium ions and high m/z range peaks The appearance of oxonium ions in the low-m/z range of the spectrum is a key component in the identification of glycopeptide spectra (Fig. 3.3). Commonly seen oxonium ions are listed below in Table 3.2.2. As reported by Carr et al., the observation of some oxonium ions is more common than others. However, all glycopeptide spectra contain the HexNAc$^+$ ion (m/z 204) and most contain a HexNAcHex$^+$ ion (m/z 366). It is also common to observe a ladder of ox-

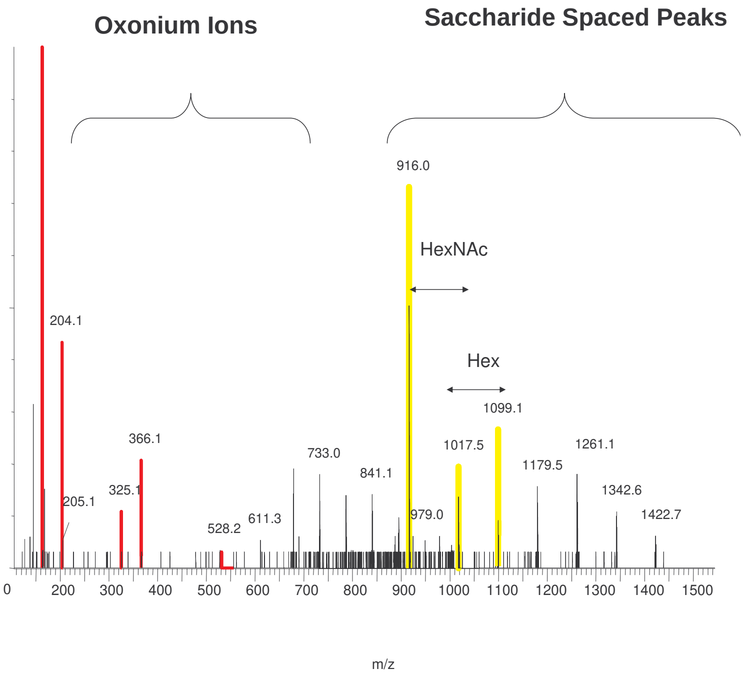


Figure 3.3: A typical glycopeptide spectrum. In this spectrum, the three main features of glycopeptide ESI-MS/MS spectra are illustrated. In the low m/z range, several oxonium ion peaks such as m/z 204 (HexNAc) and 366 (HexNAcHex) are observed. In addition, differential peak densities are observed throughout the spectrum; an area of low peak density is observed in the mid-range of the spectrum. Peaks separated by various monosaccharide combinations are also illustrated in the spectrum.

onium ions in the low m/z range of the spectrum. In figure 3.3 for example, there are oxonium ions at m/z 204 (HexNAc) and m/z 366 (HexNAcHex) as well as one at m/z 528 which represents a combined fragment of 204 and 366.

Saccharide Composition	Oxonium Ion Mass
Hexose (Hex)	162.053
N-acetylhexosamine (HexNAc)	203.079
Deoxyhexose (dHex)	146.058
N-acetyl neuraminic acid or Sialic Acid (NeuAc)	291.096
HexNac-Hex	365.132
Hex_2	324.106
HexNAc_2	406.159
HexNAc-Hex_2	527.185
HexNAc-Hex-NeuAc	656.228

Table 3.2.2 Commonly seen monosaccharides in mammalian N-linked glycans.

In addition to oxonium ions, the partially fragmented glycopeptides resulting from glycosidic bond breakage are also recorded in the high m/z range of the spectrum (see fig.3.1). Each representative peak is separated by some combination of saccharide masses. By observing the differences between these peaks in the high m/z range, the structure of the glycan can be reconstructed. In figure 3.3, the various saccharide spacings between peaks in the high m/z range are indicated.

Differential Peak Density Pattern Unlike peptide spectra, the distribution of peaks in glycopeptide spectra is non-uniform as seen in figure 3.2. Since the peptide backbone does not fragment, the oxonium ions and the partial glycopeptide fragments are separated by a mass equivalent to the unfragmented backbone. In the range representing the unfragmented backbone, generally the mid-range of the spectrum, there are very few peaks (Fig. 3.3).

In the high m/z range, the peak density is generally quite high as there are peaks representing each partial fragment with a unique mass to charge ratio. In the low m/z range, the peaks are generally quite sparse as well with the exception of the oxonium ions peaks (Fig. 3.3). This pattern of differential peak density is also a key feature of glycopeptide spectra.

3.2.3 Variations in glycopeptide spectra

The previous section described some of the main characteristics of glycopeptide spectra. However, each of these features can appear in the spectrum to varying degrees and as such our model for glycopeptide spectra should be flexible enough to accommodate all variations.

Glycopeptide spectra can vary for several reasons:

- Glycan Composition and Structure - The presence of some monosaccharides such as sialic acid as well as the structure of the glycan can bias the fragmentation of the glycopeptide. As a result, glycan structure and composition can affect the quality of the spectrum in terms of number, density and intensities of the peaks represented.

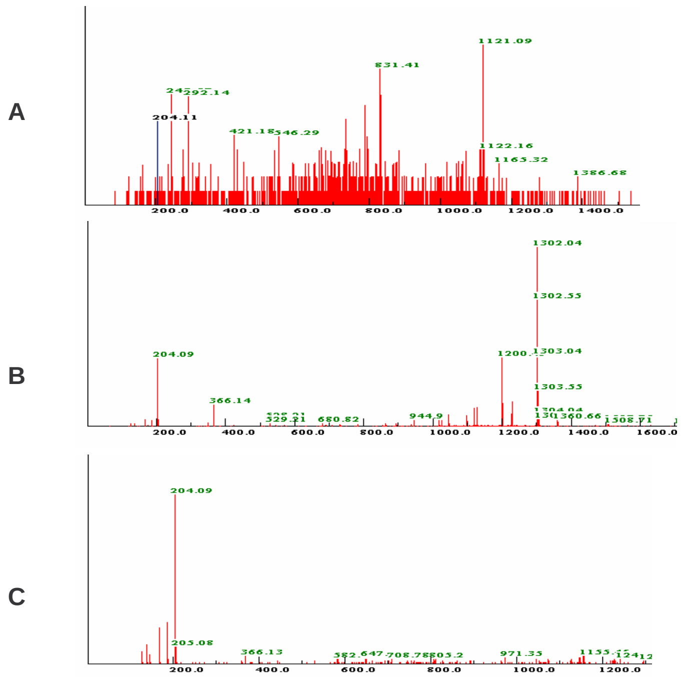


Figure 3.4: This figure demonstrates the variation of glycopeptide spectra that are observed. Part A illustrates a very noisy glycopeptide spectrum with a great deal of background noise. B shows an ideal glycopeptide spectrum in which all the attributes of glycopeptides are clearly visible. C illustrates a spectrum which is an average quality spectrum which contains several oxonium ion peaks but with only a few low intensity peaks in the high m/z range.

- Sample preparation and complexity. With increased sample complexity, in terms of concentration and number of glycopeptides, there is a greater risk of glycopeptides co-eluting and simultaneously entering the mass spectrometer. If two different glycopeptides or peptides enter at once, the resulting MS/MS spectrum can contain peaks from both species. This problem can be resolved by changing some experimental factors such as increasing the time gradient in the chromatographic steps or using different sample preparation protocols.

- Parameters utilized in MS/MS acquisition - In addition to the samples prepared for MS/MS, factors such as instrumentation can affect the quality of the acquired MS/MS. For example, peak intensities can vary with different mass spectrometers which have varying sensitivities. As well, factors such as collision energy, matrix, charge state, and the type of ion formed will cause glycopeptides to fragment to varying degrees and thus affect the quality of the MS/MS spectrum[20].

3.2.4 Relative Importance of Each Feature to the Identification of Glycopeptides

The presence of oxonium ions, partial glycan fragments and different peak densities are the main features utilized by glycoprotein chemists in the identification of glycopeptides. However, it is quite common to observe each of the features individually in non-glycopeptide spectra and as such, a combination of each of these features is necessary for the glycopeptide classifier to be effective.

The discriminating ability of each feature was assessed and used to assign the relative importance of each feature. The following weights were assigned to each feature:

- 50% - Oxonium Ion Presence. The presence of peaks located at known oxonium ion m/z values is the most informative feature in glycopeptide detection. Oxonium ion masses however are not completely unique[1]. In figure 3.5, a peak with m/z 204.13 is highlighted. Although this mass coincides with that of a HexNAc oxonium ion[30], this peak represents a peptidic GK fragment. Thus, the presence of oxonium ions alone is often insufficient for the identification of glycopeptides.

- 40% - Differential Peak Distribution. Although non-glycopeptide spectra have mainly uniformly distributed peaks, it is likely that peak densities can randomly vary in these spectra. Peak density alone therefore is not a reliable metric for glycopeptide identification.

- 10% - Saccharide-spaced peaks in the high m/z range. It is highly likely that peaks appearing in MS/MS spectra are separated by mass differences equal to various combinations of saccharides by chance alone. For this reason, it is given a weight of only 10%. However, these spacings may not appear contiguously and may not possess the correct charge, and thus saccharide-spaced peaks should still be included in the glycopeptide model.

[1] Oxonium ions have unique masses when given a high enough precision. For example, a HexNAc oxonium ion has a precise mass of 204.09 whereas a peptidic y2-GK fragment has mass 204.13. However since there is a limitation on the precision of the mass spectrometer (0.1 daltons), there may be ambiguity in the assignment of this peaks.

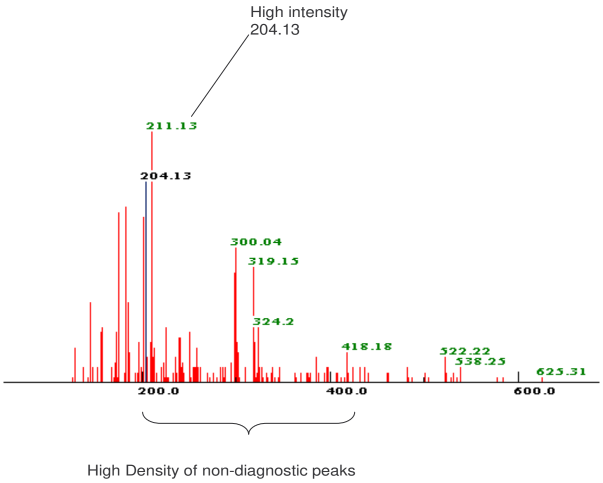


Figure 3.5: This glycopeptide spectrum contains a high intensity peak at mz/ 204.13, the same mass as a HexNAc oxonium ion fragment. However, this spectrum represents a peptide. The 204.13 peak in this case represents a y2-tryptic fragment of GK di-peptide.

3.3 Functions for Glycopeptide Attribute Evaluation

In this section, we will describe the functions used for evaluating each of the attributes of glycopeptide spectra described in the previous section. Given an input spectrum, the goal is to develop a function f for each feature which returns a score reflecting the expression of the particular attribute in an experimental spectrum. Standard mass spectrometers produce output as a vector of pairs of real numbers (mz, intensity). Each function f therefore takes in as input vector E which represents all (mz, intensity) pairs of the experimental spectrum.

Each f_i for attribute X_i was derived such that, based on the weights assigned to each feature as described in the previous section, w_i, the sum of each $\mathrm{w}_i\mathrm{f}_i$ for an idealized glycopeptide spectrum can be normalized to a score of 1. A total score S for glycopeptide classification can thus be described as:

$$S = f_{OxoniumIons} * 0.5 + f_{peakdensity} * 0.4 + f_{partialglycanfragments} * 0.1 \quad (3.1)$$

Given the variations of glycopeptide spectra discussed in section 3.2.3, each f_i developed should be sensitive enough to assign a correct score to noisy glycopeptide spectra while being discriminating enough to eliminate false positives.

3.3.1 Oxonium Ions Attribute Function

An attribute function f to evaluate the presence of oxonium ions in an MS/MS spectrum should return a value indicating our confidence that the appearance of peaks at the m/z values of oxonium ions is not random. To accomplish this requires the incorporation of information other than just peak m/z values.

Peak Significance Measures One of most important criteria in assessing the validity of a peak in an MS/MS spectrum is its intensity. In general, there is a higher probability that a more intense peak represents a valid fragment. However, peak intensity also depends strongly on the physical and chemical properties of the glycopeptides, so it is often incorrect to assume that intense peaks are more valid than the weaker ones. In carbohydrate spectra, peaks with low intensity often represent valid fragment structures, but which due to the chemical property of the glycan, are less likely to ionize.

Since ESI-MS/MS spectra exhibit a great deal of random noise, in some spectra almost at every m/z unit[33], during the processing of the data the mass spectrometer determines the background noise level and normalizes all peaks of the spectrum according to this value. A common metric used to distinguish a valid peak from background noise is that the peak should be at least 3 times as intense as the background noise level.

Multiple Oxonium Ions When several oxonium ions are found in the spectrum, there is added confidence that the occurrence of each one is not a random event. Further, as discussed in section 3.2.2 oxonium ions can form a ladder of peaks. If multiple oxonium ions form a ladder of peaks, there is a higher probability that the peaks are not random. For example, if significant oxonium ions of 204 (HexNAc) and 366 (HexNAc-Hex) are both observed in addition to a peak at m/z 528 representing (HexNAc2-Hex), the presence of all 3 peaks simultaneously increases the probability that each individual peak represents a valid oxonium ion.

Peak Density In ideal glycopeptide spectra, fragments found in the low m/z range should consist of only oxonium ions peaks (Fig. 3.3) since the peptide backbone does not fragment. As such, the ratio of diagnostic peaks to non-diagnostic peaks in this m/z range should be fairly high.

The density of peaks which do not represent oxonium ions is an additional metric which can assess the validity of the entire set of oxonium ions observed in the spectrum. In the example illustrated in figure 3.5, the density of peaks surrounding the peak at m/z 204.13 suggests that the spectrum does not represent a glycopeptide. If the set of all oxonium ion peaks are among the most intense peaks in the low m/z range of a spectrum, there is additional confidence that the oxonium ion peaks are valid.

Function for Oxonium Ions Evaluation The definition of a function evaluating oxonium ion content incorporates a sum representing the validity of each oxonium ion found in the spectrum and also an evaluation of the set of all oxonium ion peaks found in the spectrum.

To evaluate the validity of the oxonium ion peaks, a score was derived based on 3 factors:

- The sum of all the intensities of significant oxonium ions found in the spectrum
- A constant factor α evaluating the presence of the oxonium ion in glycopeptide spectra. Weights based on the probability of observing a specific oxonium ion were assigned and these values incorporated into the score.

- A constant factor of β which evaluates the presence of an oxonium ion ladder. If a peak representing a di- or tri-saccharide oxonium ion is observed along with its component monosaccharides, a constant factor of β is added to the score.

To incorporate information about the entire set of oxonium ions found, a metric δ was derived to evaluate the ratio of non-oxonium ion peaks to oxonium ion peaks in the low m/z range. This score was subtracted from the peak validity score to penalize very dense spectra which randomly contain peaks at oxonium ion m/z values.

Overall, the function for oxonium ion evaluation can be defined as follows:

$$f_{OxoniumIons} = (\sum_{j=1}^{m} (\alpha_j + \beta_j) * Intensity(j)) - \delta$$

where m is the total number of significant oxonium ions detected in the input spectrum.

3.3.2 Differential Peak Densities

The second most important feature, assigned a weight of 40% in the glycopeptide model (eqn.3.1), is the observation of a pattern of differential peak density in the spectrum. The high m/z range peak density was not considered as it was found that many of the spectra obtained were of low quality and often did not contain many peaks contributing to an inflated false positive rate.

To derive a measure of the sparsity of the low and mid-range m/z ranges, a tally of the significant peaks which do not represent known oxonium ions was taken in each range. The number was then discretized to a score out of 40 to represent 3 qualitative classifications of peak densities: sparse, not sparse, and dense.

3.3.3 Monosaccharide Loss

Although most glycopeptide spectra can be identified by the presence of oxonium ions and differential peak density, an additional metric used in the identification of glycopeptide spectra is the presence of peaks separated by combinations of monosaccharides (eqn.3.1). A simple function to assess this feature in experimental spectra was a tally of the number of peaks separated by masses of 203 (HexNAc) or 162 (Hex) in the high m/z range.

3.4 Score Significance

Once the score for each experimental spectrum is obtained based on equation 3.1, the score must be mapped to either represent a glycopeptide spectrum or a non-glycopeptide spectrum. To classify a spectrum as belonging to a glycopeptide requires the establishment of a decision score. The selection of an appropriate decision score is discussed further in section 5.4.

Chapter 4

Glycan Analysis Module

As described in chapter 1, protein glycosylation can drastically alter protein function and structure. Each cell type produces glycans with specific structures and monosaccharide compositions, which in turn affect the function of the glycoprotein. Given the close link between glycan structure and function, structure elucidation is an important component of glycoprotein analysis.

4.1 MS/MS For Glycan Analysis

Limitations of MS/MS analysis for glycan structure elucidation Traditional methods of glycan structure determination such as methylation analysis or exoglycosidase digestion[30] often require large amounts of sample and can not be processed in a high throughput manner. As previously mentioned, mass spectrometry is increasingly a popular choice as it can operate on very small amounts of sample[30] (see section 1.3.5). Despite the popularity of mass spectrometry in glycan structure analysis, there are several limitations of utilizing this technique. For one, mass spectrometry does not provide information about bond anomericity[30], a description of the type of glycosidic bond (α or β) between each monosaccharide residue, that can affect the chemical properties and function of the glycan[30]. In addition, MS cannot distinguish between monosaccharides with the same mass, e.g. hexoses glucose, mannose, galactose (mass 162); or hexosamines glucosamine, galactosamine (mass 204)[6]. To obtain the exact structure of glycans therefore, a combination of mass spectrometry and traditional carbohydrate analysis methods must be used. Despite these limitations of MS/MS in glycan structure elucidation, it can still provide valuable structural information.

Manual Analysis of Glycan Spectra The reconstruction of glycan structures from glycopeptide MS/MS spectra is complex and labor-intensive. Manual glycan structure determination involves detecting mass differences between the high intensity peaks of the spectrum. The order in which the mass differences are observed between the various peaks suggests the order of monosaccharide dissociation and thus the composition of the glycan. Multiple monosaccharide differences originating from the same peak and the relative intensities of the peaks observed also suggests the branching points in the glycan. With the incorporation of known rules about glycan structure and biosynthesis, the branch points and the monosaccharide composition, the glycan structure can be elucidated. Obfuscating factors such as missing or additional peaks and multiply charged peaks in ESI-MS/MS however, can complicate the task of glycan structure determination significantly.

4.2 Exisiting Methods for Automated Glycan Analysis

Many existing methods for MS/MS spectra interpretation described in section 2.2.1 deal primarily with peptide sequencing and are not directly applicable to the analysis of glycans. To date, there have been no techniques developed for the automated glycan analysis from glycopeptide spectra, and only a few methods for automated glycan analysis.

One of the earliest tools developed is the Saccharide Topology Analysis Tool (STAT) developed by Gaucher et al. [12]. STAT is a web-based tool that can extract glycan sequence information from a set of MS/MS spectra for an oligosaccharide of up to 10 residues. Given information such as precursor ion mass, possible monosaccharide moieties, charge carrier, and product ion mass from the user, all possible structures are generated and evaluated against experimental glycan spectra. The list of possible structures is given a rating based on the likelihood that it is the correct sequence in accordance with glycan biosynthetic rules and presented to the user.

Mizuno et al.,[22] have developed an automated program which assigns known losses of monosaccharides to peaks in Post-Source Decay (PSD) spectra of N-linked glycans. They have also developed a spectrum simulator to generate hypothetical tandem mass spectra. The comparison is not automated and users must generate all structures they think are possible and compare them manually. StrOligo developed by Ethier et al., [11] is another tool for automated complex-type N-linked glycan analysis. StrOligo consists of two main modules. The first one deconvolutes the MS/MS spectra and creates a singly-charged peak representing multiply charged peaks to facilitate the process of glycan structure determination. It then analyzes the deconvoluted spectra to search for mono- and di-saccharide masses found between the major peaks and creates a relationship tree between all the masses found. The possible monosaccharide combinations for each peak in the relationship tree is calculated and the most likely compositions presented to the user who then selects the most likely structure. StrOligo was found to be capable of determining the correct structure in 86% of the glycans analyzed and produced the top three top scoring results in 100% of the glycans analyzed.

More recently, Lohmann et al [20] put forth Glyco-Fragment, a tool intended to support the manual assignment of all peaks contained in the mass spectra of complex carbohydrates. Glyco-Fragment is a web-based tool that reads in a glycan structure, determines the molecular structure and generates all possible fragments. However, there is no methodology to automatically assign these fragment peaks to MS/MS spectra.

The automated methods described above are not all applicable to the analysis of ESI-MS/MS glycopeptides. For one, all methods described above were developed for the analysis of derivatized oligosaccharide spectra and not glycopeptides. In addition, the methods of Ethier, Gaucher and Mizuno are similar to *de novo* spectra analysis methods. Since glycopeptides in complex mixtures do not ionize well in proteomic data[20], the resulting spectra often contain missing peaks and may be of low quality and as such inappropriate for *de novo* techniques. Another limitation of existing methods for automated glycan analysis is that they require input from the user regarding the possible composition and structures of the glycans analyzed in the sample. In a high throughput environment however, obtaining this input may not be feasible for reasons such as the high volume of data produced or the user not having any insight into the types of glycans to be expected in the sample.

Given the drawbacks of the existing methods for automated glycan analysis, we present a methodology for automated glycan structure determination by adapting MS/MS ion searching techniques for glycan analysis.

4.3 MS/MS Ion Searching Techniques as applied to glycopeptides

To automate the process of glycan structure elucidation from ESI-MS/MS glycopeptide spectra, we present an approach based on the adaptation of traditional techniques of MS/MS ion searching for glycan analysis (see section 2.2.1).

Most MS/MS ion searching techniques thus far have catered to peptide fragmentation and are not applicable to glycopeptide analysis. For application to glycan analysis, existing peptide MS/MS ion searching techniques need to be modified in two main respects:

- The branched structure of carbohydrates requires a unique model for theoretical fragmentation (section 4.5).
- The unique features of glycopeptide spectra require a novel method for spectra correlation (section 4.8).

As in traditional MS/MS ion searching, our approach for glycan ion MS/MS ion searching involves three main steps:

1. Obtaining a suitable database of structures which could correlate to the experimental spectra

2. Generating theoretical spectra representing predicted fragmentation products of each of the database entries

3. Correlating the theoretical spectra to the experimental spectra and determining the most likely match.

Each of these steps will be further discussed in the following sections.

4.4 Glycan Database

A total of 4469 N-linked glycan structures were obtained from GlycoSuite DB (Proteome Systems Limited) [1]. GlycoSuiteDB is a relational database that curates information from scientific literature on glycoprotein-derived glycan strucures, their biological sources and the references in which the glycan was described [6].

The glycans obtained from the database do not provide a complete set of all N-glycans found in nature and it is possible that not all experimental glycan spectra match exactly with database glycans. As discussed in section 2.2.1, the reliance of MS/MS ion searching techniques on the completeness of the database used is an inherent limitation of the technique. However, a secondary goal of MS/MS ion searching techniques is to return the most similar or homologous structure in the case where the experimental structure is not reported in the database. Since N-linked glycans have a well-defined structure and are generated by similar biosynthentic mechanisms, it is likely that the database will contain a similar glycan in case the exact structure is not found in the database.

[1] Obtained in July 2002 with an academic license.

4.5 Algorithm for Glycan Carbohydrate Fragmentation

4.5.1 Characteristics of Glycan MS/MS Spectra

Unlike peptide fragmentation, carbohydrate fragmentation is complex due to the presence of branches. Theoretical peptide fragments are created by breaking each of the peptide bonds and adding the masses of the amino acids of the resulting fragments. The number of partial fragments created will in theory equal the number of peptide bonds present[2] (Fig. 4.1a). Glycan fragmentation however is much more complex since there can be simultaneous glycan fragmentatation along each branch of the glycan structure. As such, the set of peaks produced will include some peaks representing combinations of masses between partially fragmented branches (Fig. 4.2).

The number of fragments observed in carbohydrate spectra however, is much smaller than the set of all predicted fragments. For one, not all fragment species are produced with the same probability. The structure and composition of each carbohydrate introduces a bias for the observation of some fragmentation products more than others[12]. The chemical properties of individual monosaccharides can also introduce a fragmentation bias. The weaker bond energy of sialic acid residues for example causes them to dissociate more readily than other monosaccharides. Another factor influencing the number of glycan fragments observed is the energy of dissociation used for the fragmentation. High energy collisions will break more glycosidic bonds in the structure and as such contribute to the observation of more fragment species and more peaks in the spectrum. In the concomitant analy-

[2] Considering either the b-, or y- ion series

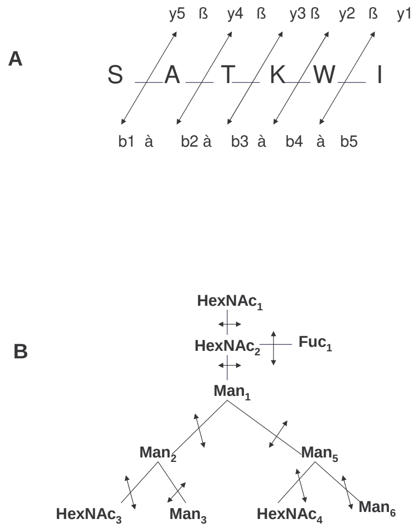


Figure 4.1: This figure illustrates the fundamental difference between peptide and carbohydrate fragmentation. Potential fragmentation points are illustrated with double-ended arrows. A) The linear peptide molecule fragments at the peptide bonds and creates b- or y-type ions. Peptides have as many possible breakage points as there are residues and for any one type of fragment product (i.e. b- vs. y-ions), the number of peaks produced is at most the same as the number of bonds. The branched structure of carbohydrates as illustrated in B however, has potential fragmentation points all along the structure. Since there are 2 branches for the structure in B, there can be 2 simulataneous fragmentation events, one along each branch resulting in a bigger set of possible peaks.

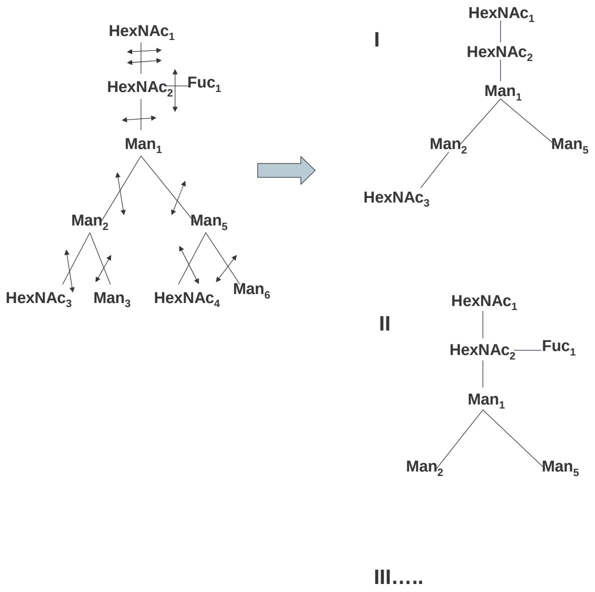


Figure 4.2: The number of fragments derived from carbohydrate CID can be quite large due to the need to consider fragmentation products across branches. In this schematic, two CID species are illustrated. Species I and II represent unique masses generated by partial fragmentation across the two branches. Thus, in addition to having to consider fragmentation products along each path, sub-tree combinations must also be examined.

sis of glycopeptides and peptides, the energy of dissociation may not enable the generation of all possible glycan fragments and as such only a subset of all likely fragments are observed [3].

Another major reason that the number of observed peaks is much smaller than all possible peaks is that many fragmentation products have the same composition and thus the same m/z ratio. Carbohydrates in higher animals are composed of a maximum of 6 monosaccharides of which two are rare (Table 4.5.1). As such, for any glycan it is likely that various fragments contain the same monosaccharide composition and thus produce only 1 peak in the resulting MS/MS spectrum. and identical peak masses.

Monosaccharide	Mass
Hexose (galactose, glucose, mannose)	162.053
Hexosamine (GlcNAc, GalNAc)	203.079
Deoxyhexose (Fuc)	146.058
Sialic Acid (NeuAc)	291.096
Pentoses (Xyl)	132.042
Uronic Acid (GlcA, IdA)	176.032

Table 4.5.1 Commonly seen oxonium ion peaks in glycopeptide spectra.

[3] The energy of dissociation used during MS/MS acquisitions is determined by a scaling function based on the mass of the species entering the MS/MS chamber. The function is applied to a specific mass range which may be directed towards peptide analysis. Often, glycopeptides surpass the upper peptide mass limit and when they enter the MS/MS chamber and the energy determined by the scaling function, that of the upper limit of the peptide mass range, is inadequate to produce all possible fragments.

4.5.2 The Full Model of Glycan Fragmentation

Without a complex model for carbohydrate fragmentation, there is no simple method to predict the glycan fragments produced by CID. A simple approach to this problem is to assume that all glycosidic bonds in the carbohydrate structure are equally susceptible to fragmentation and consider all possible fragments. This model, the Full Model of Glycan Fragmentation, while not realistic, provides a complete set of fragments.

One of the major drawbacks of this method is that the size of the set of theoretical fragments generated can get very large. We can enumerate the number of possible fragments that can be created by the Full Model of glycan fragmentation by considering N-linked glycans as rooted, binary trees. N-linked carbohydrate structures found in nature all contain the pentasaccharide core $HexNAc_2Man_3$ from which stems two antennae, or branches. There are several tri-antennary structures although they are not as common as bi-antennary structures[4]. Bi-antennary glycans therefore assume rooted, binary tree structures with nodes representing the monosaccharide residues and a root representing the initial $HexNAc_2Man$ portion of the N-linked core (see structures illustrated in Fig. 4.4).

Let F be a set of all the possible unique masses that can be obtained from the full fragmentation of a carbohydrate structure C. $\mid F \mid$ can vary depending on four main factors. The effects of each of these factors on $\mid F \mid$ is illustrated in figure 4.3.

- **Glycan Size**. The number of nodes, which in biological terms translates to the mass of the glycan, is directly proportional to the size of

[4] There are also some N-linked glycans with a single GlcNAc residue, called a bisecting GlcNAc, attached to the core inaddition to the two antenna. These structures are rare compared to N-linked glycans.

the F (Fig.4.3). A simple example of this is illustrated in figure 4.3. The smaller glycan in 4.3b will produce less fragments that the glycan in figure 4.3a. Even the addition of one extra residue, depending on where it is added, can significantly increase the number of peaks produced in the Full Model. In figure 4.3d, the effect of an extra core Fucose residue increases the total fragmentation products by a factor of 2 compared to the glycan in figure 4.3a.

- **Glycan Topology** For a carbohydrate structure with n nodes, there are a total of $(2n-3)!!$ different rooted, binary, tree topologies[9]. The topology of the glycan can affect the number of peaks produced by the Full Model. For a glycan with m branches, there can be any number of fragmentation events from 1 to m since there is a maximum of 1 fragmentation event per branch. Thus, the topology directly affects the number of possible branches, which in turn influences the number of different mass combinations possible with full fragmentation. For example, for two glycans with identical composition $\mathrm{HexNAc_5Man_4}$, the caterpillar topology illustrated in figure 4.4a produces more fragmentation products compared to the more full glycan structure illustrated in figure 4.4b in which there are many redundant masses produced. In figure 4.4a there can be a maximum of two fragmentation events as no two fragmentations can occur along the right branch of the structure.

- **Monosaccharide Composition.** Since the N-linked glycans of mammals usually comprise of 6 unique monosaccharide masses [30], for each topology there are a maximum of 6^n different monosaccharide combinations for n residues. In reality however, this number will be much less as only some monosaccharide arrangements are biologically valid. In

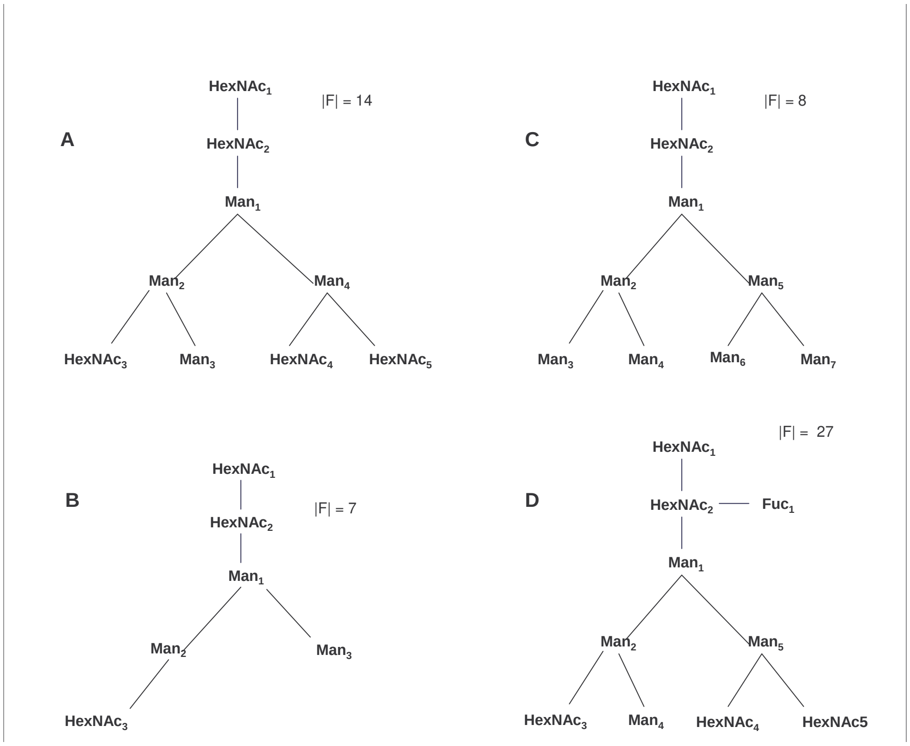


Figure 4.3: Effects of glycan size and monosaccharide composition on | F |. In figure A we observe a glycan which under the full fragmentation model produces 14 unique peak masses. When a similar structure with a reduced number of nodes as illustrated in B, a less massive and complex glycan, is fragmented with the Full model, |F| is reduced to 7. Figures C and D demonstrate the effect of monosaccharide composition on |F|. In C, a high mannose glycan with less monosaccharide variability but with similar topology to A, produces a smaller set of peaks under the Full Model than A. Figure D illustrates the effect of an additional branch, the core Fucose branch attached to HexNAc$_2$ which also increases the size of F.

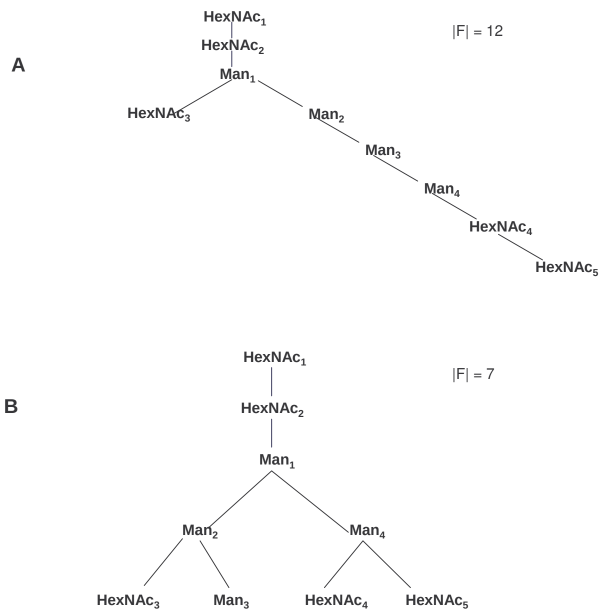


Figure 4.4: Effects of glycan structure on the number of peaks produced by the Full Model. This schematic illustrates the difference of | F | between two glycans with the identical monosaccharide composition, HexNAc$_5$Man$_4$. In A) we observe a glycan with a caterpillar type topology which produces a larger number of unique peak masses compared with the glycan with topology illustrated in B. The combinations of branches in B produces redundant masses and thus the set F is smaller.

addition, since various combinations of monosaccharide will produce identical masses, the actual number of possible peaks will be reduced. In figure 4.3c, the effects of monosaccharide composition can be seen. The leaves in figure 4.3c are all mannose as opposed to the leaves of figure 4.3a which have a varied monosaccharide composition. The additional HexNAc molecules in figure 4.3a, increase the total number of unique mass combinations possible.

4.6 Evaluation of the Predictive Power of the Full Model

The Full Model of glycan fragmentation produces the entire set of theoretical glycan fragment peaks regardless of the likelihood of observing a particular fragment. As previously mentioned, due to the chemical structure of glycans, only a subset of all possible peaks is actually observed. Although the set of peaks produced by the Full Model is complete, when the theoretical fragment peaks are correlated to experimental spectra, the unlikely peak fragments can be matched to random or noise peaks. This problem is illustrated in the example in 4.5, in which an unlikely fragmentation product of the Full Model is incorrectly matched to a peak in the experimental spectrum. Figure 4.5a illustrates a glycan prior to fragmentation. Upon CID, the highly labile terminal sialic acid residues of 4.5a will most likely dissociate before G1, G2 and G3 as indicated in 4.5b. When the Full Model of fragmentation is applied to 4.5a however, unlikely peaks such as those illustrated in 4.5b will be produced. Given that in some spectra, there is a peak at almost every m/z unit, it is possible that these unlikely fragments are incorrectly matched.

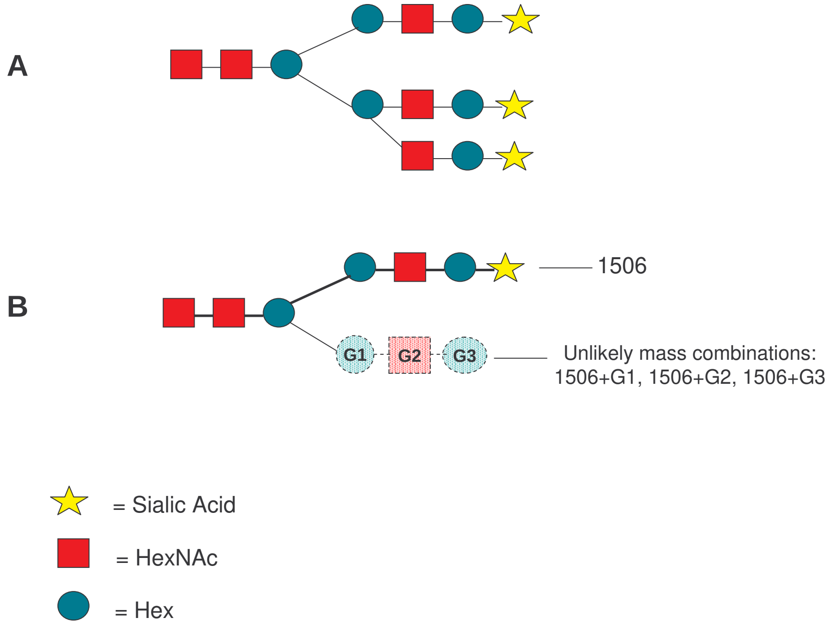


Figure 4.5: This figure illustrates one drawback of the Full Model of Glycan Fragmentation, namely the generation of unlikely fragmentation products. In A, we see a glycan with composition $\mathrm{HexNAc_5Man_5NeuAc_3}$. Since sialic acid is a highly labile molecule, fragments such as those observed in B, will not be observed since the terminal sialic acid will dissociate before residues G1, G2 and G3 as indicated. However, these unlikely peaks will be generated in a theoretical fragmentation of A using the Full Model.

Another potential problem with the Full Model is that it may not scale well with increased glycan size. Unless an appropriate scoring scheme is devised for the correlation of the theoretical and experimental spectra, larger glycans will create a larger set of peaks and will be matched more frequently to random peaks in the experimental spectrum.

4.6.1 The Path Model

As discussed in the previous section, the set of peaks created by the Full Model of carbohydrate fragmentation can become very large depending on the structure and size of the glycan and consequently lead to increased false peak matches. In this section, we present an alternate fragmentation model which produces a set of peaks $S \subseteq F$ that are complete enough to correctly match database glycan structures to experimental glycan spectra but which do not produce a great number of unlikely fragment peaks.

In a previous study, Mizuno et al found that ions produced by single-bond cleavages were more abundant than fragment ions resulting from multiple-bond cleavages[22]. Further, they concluded that fragmentation initiated in a branch proceeds to the end of the same branch[22]. Based on this result, the Path Model of glycan fragmentation was developed. In this model, to obtain all possible fragmentation products resulting from fragmentation along one branch, an in-order traversal of the carbohydrate structure is performed. The process of glycan analysis using the Path Model of fragmentation is illustrated in figure 4.6.

One of the main advantages of the Path Model over the Full Model number is that the number of fragments produced is proportional to the number of monosaccharides in the structure regardless of the size and topology of

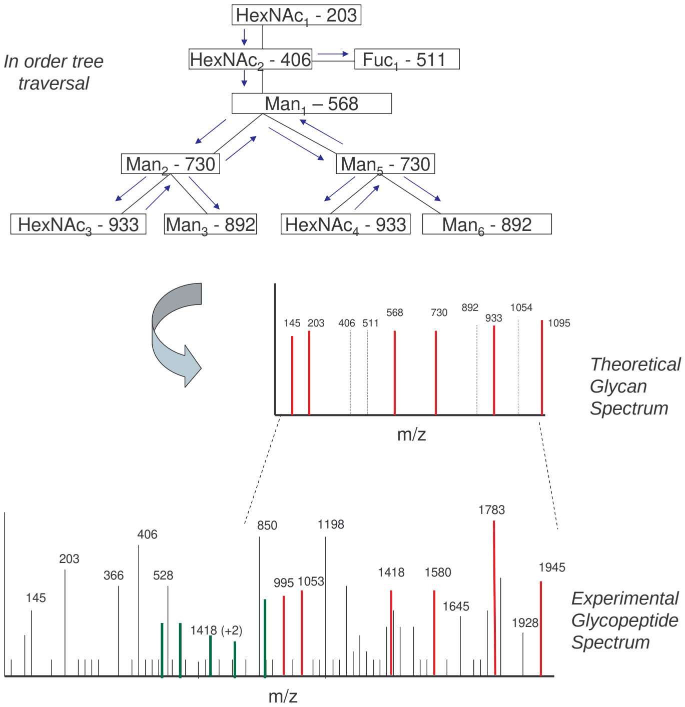


Figure 4.6: This figure illustrates glycan MS/MS ion searching using the Path Model of glycan fragmentation. Peaks generated by the and in-order traversal of the glycan structure (the Path Model) are overlaid on an experimental spectrum and correlated.

the glycan. As such, the likelihood of false peak matches, especially in larger glycans, may be smaller compared to those produced by the Full Model. The Path Model of glycan fragmentation was implemented and it's effectiveness examined in section 5.7.

4.7 Evaluation of the Predictive Power of the Path Model

The Path Model for glycan fragmentation provides a solution to many of the problems posed by the Full Model. However, in some situations, the Path Model may be inadequate in returning a correct glycan structure. Although the appearance of single-bond cleavages are more abundant than multiple-bond cleavages[22], there may still be several peaks present representing multiple-bond breakages. Thus, a peak in the experimental spectrum produced by the combination of monsaccharides between branches may be matched to a structure that is longer along one branch. For example, in figure 4.7, the correct glycan structure (Fig.4.7a) is not returned as the peak representing the combined mass between monosaccharides 4 and 5 is not produced. Instead, the mass between 4 and 5 is produced by an incorrect linear structure illustrated in figure 4.7b. When an MS/MS spectrum representing the correct structure is interpreted manually, factors such as the peak intensities and known rules about glycan structure would suggest the actual structure of the spectrum is that of figure 4.7a and not figure 4.7b.

4.8 Algorithm for Spectra Correlation

After the creation of the theoretical spectrum modelling carbohydrate fragmentation, the theoretical spectrum is correlated with the experimental spectrum in order to identify the correct glycan structure. In theory, the best correlation will identify the glycan represented in the experimental spectrum. Glycopeptide spectra correlation differs from existing methods for peptide spectrum correlation in two main ways:

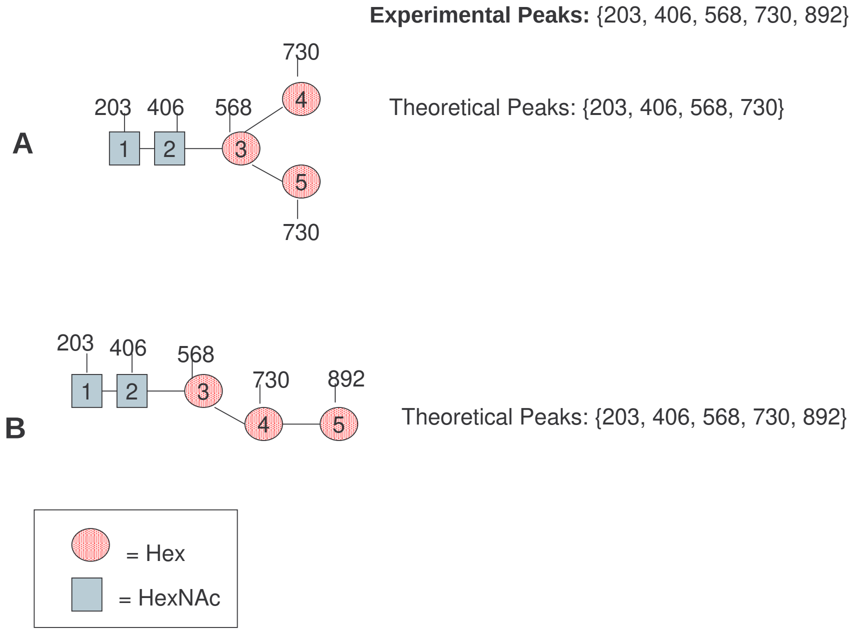

Experimental Peaks: {203, 406, 568, 730, 892}
A
203
406
568
730
730
1
2
3
4
5
Theoretical Peaks: {203, 406, 568, 730}
B
203
406
568
730
892
1
2
3
4
5
Theoretical Peaks: {203, 406, 568, 730, 892}
= Hex
= HexNAc

Figure 4.7: This figure illustrates a potential problem with the Path Model. The glycan in A is the correct structure matching the experimental spectrum. However, fragment 892, a combined mass fragment between residues 3,4 and 5, found in the experimental spectrum is not considered in the Path Model. As such, when the Path Model is applied to the glycan in B which contains fragment 892 as an extra Hex residue along one branch, this structure is returned as the correct glycan.

- **Unknown point of attachment of the glycan to the peptide backbone.** Since the peptide moiety of glycopeptides remains intact after fragmentation, the peak representing the starting point of the glycan is not immediately known. When analyzed manually, this peak, the naked peptide peak, is determined by tracing monosaccharide loss sequentially and finding the most likely point of attachment.
- **Detecting branching patterns in the spectra.** The peaks created by theoretical fragmentation must also be correlated to the structure of the glycan. It is possible that the set of theoretical fragments derived from a glycan with an incorrect structure but similar composition be falsely matched to an experimental spectrum. As such, glycan structure should be taken into account when deriving an appropriate scoring scheme to evaluate the degree of matching between the theoretical and experimental spectra.

In the following sections we describe the approaches used in the correlation of the theoretical spectra and the experimental spectra of glycopeptides.

4.8.1 Naked Peptide Determination

In order to match the theoretical glycan peaks of the experimental spectrum, we need to determine the offset of the peak representing the peptide moiety, the 'naked peptide' in the experimental spectrum. Since the naked peptide peak of the glycopeptide is not always easily identifiable, it is necessary to determine this point before the correlation of the spectra can begin.

In N-linked glycopeptide MS/MS spectra, the naked peptide peak with an attached GlcNAc residue is typically amongst the most intense peaks of the spectrum. A simple approach to determining the naked peptide is to generate a list of the most intense peaks in the high m/z range of the spectrum, and try each one[5] as a potential starting point. In theory, when the correct database glycan is applied on the spectrum at the correct point, there should a maximal number of matching peaks and thus the highest correlation score. The top hits therefore, should provide the optimal sugar structure matching the peaks as well as the most likely naked peptide.

Determining the most likely naked peptide peak also provides valuable information required in matching the glycopeptide to its parent protein. Once the exact mass of the peptide without the glycan moiety is determined, the peptide can be matched to its parent protein by the application of Peptide Mass Fingerprinting (PMF) Techniques (see section 2.2.1). This process is illustrated in figure 4.8.

4.9 Correlation of Theoretical and Experimental Spectrum

From each naked peptide candidate, the peaks of the theoretical spectra are matched to those in the experimental spectrum. To evaluate the degree of matching, an appropriate scoring scheme must be developed. As with MS/MS ion searching for peptides, the scoring scheme used in evaluating the degree of matching between the theoretical and experimental spectrum is crucial in the performance of the technique[1].

The majority of MS/MS ion searching programs incorporate 3 main features:

- **Number of matched peaks**. This metric describes the number of

[5] Since the naked peptide could be a +2 or +3 charged peak, all of the charge states of the naked peptide are also tried as potential starting points.

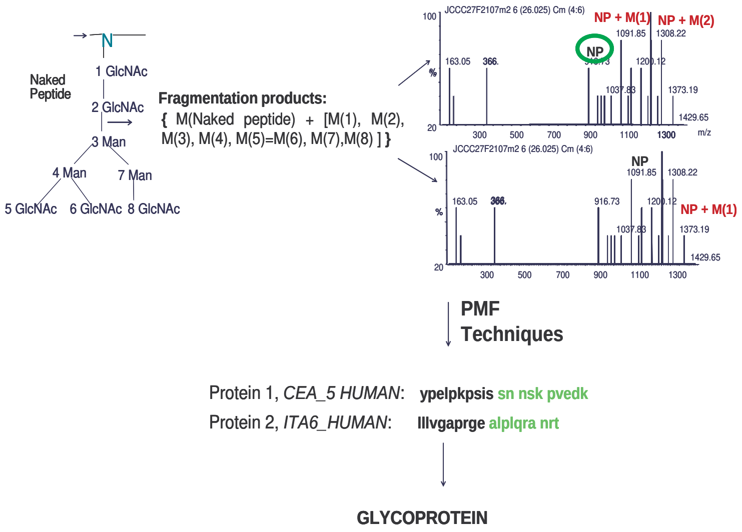


Figure 4.8: The determination of the naked peptide peak enables the matching of the glycopeptide to its parent protein. In the example illustrated in this figure, the glycan shown is fragmented using the Path Model of glycan fragmentation. These peaks are subsequently overlaid upon experimental glycopeptide spectra and scored starting from various high intensity peaks in the high m/z range, each a naked peptide candidate. From the highest scoring match, the naked peptide and glycan are determined. The naked peptide mass can then used to match the glycopeptide to its parent protein using Peptide Mass Fingerprinting (PMF) techniques (see section 2.2.1.)

theoretical peaks that are found in the experimental spectrum.

- **Completeness of sequence ladder**. In peptide spectra, if a complete ladder of b- and y-ion fragments are found in the spectrum, there is a greater probability that the peaks represent valid peptide fragments.

- **Matched Peak Intensity**. Since peak intensity represents the number of fragments found at the same mass, a higher peak intensity suggests a higher likelihood that the peak represents a valid fragmentation product as discussed in section 3.3.1. To assess the validity of all matched peaks and thus to the overall likelihood that the theoretical and experimental spectra are correlated, the intensities of all matched peaks are added to the correlation score.

In addition to these features common to most peptide MS/MS ion searching techniques, various programs incorporate other specific information in the evaluation of the matching between theoretical and experimental spectra. For example, SCOPE developed by Bafna[1] utilizes a probabilistic scoring function that incorporates detailed knowledge on how peptides fragment and some specific features of peptide spectra. Prob ID developed by Zhang et al, is another probabilistic approach which incorporates specific information about the peaks of the experimental spectrum such as the presence of immonium ions and also evaluates noise and unmatched peaks in the spectrum [33].

4.9.1 Glycan Spectra Correlation Scoring Scheme

In this section, we present a scoring scheme to evaluate the matching between the theoretical glycan spectra and experimental glycopeptide spectra.

As with the scoring schemes used in peptide MS/MS ion searching, the intensities and number of matched peaks are incorporated in the scoring scheme. In addition to these common features, it is necessary to incorporate some information on the structure of the glycan.

4.9.2 Isotope Modeling for Charge Determination

To generate the theoretical spectra to correlate to experimental glycopeptide spectra, glycans from GlycoSuite DB were fragmented (using either the Path or Full Model of glycan fragmentation) and the resulting peaks searched in the experimental spectrum. When searching a theoretical peak in the experimental spectrum, the peak masses are checked in several charge states as peaks in ESI-MS/MS spectra exist in +1, +2 and +3 charges. For this reason, a model for isotope detection was developed in order to differentiate between singly, doubly and triply charged peaks of the experimental spectrum. This process of isotope detection is a well-researched subject and there are several approaches to this problem. As an initial approach, the charge state of a peak was determined by determining the most intense peak in a given window and observing the peak spacings surrounding the peak. Spacings of 0.5 denoted doubly charged peaks, 0.33 triply charged peaks and 1.0 singly charged peaks. For a theoretical peak to match an experimental peak, the corresponding charge states of the two peaks had to match to be scored.

Glycan Structure Evaluation Experienced glycoprotein chemists can often recognize the structure of the glycan based on the overall appearance of the spectrum. In addition to information about monosaccharide composition by observing saccharide-spaced peak distances, other factors such as relative peak intensities suggest the correct glycan structure. Oligomannose-type

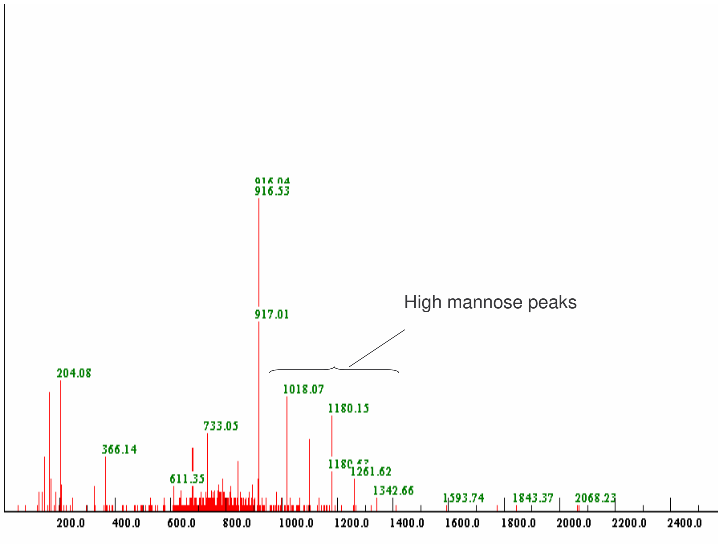


Figure 4.9: This spectrum represents an oligomannose glycopeptide. The appearance of evenly spaced, intense peaks indicated in the figure suggests the glycan type is high-mannose.

glycan spectra for example are easily identifiable by observing a series of evenly spaced high intensity peaks (Fig.4.9.). Thus, determining the correct structure of a glycan from MS/MS spectra in an automated fashion requires thorough modelling of glycan fragmentation and the relative peak intensities expected from fragmentation, a complex task.

As an alternative to verifying the structure of the entire glycan, we propose to examine glycan substructures. To achieve this task, the theoretical fragments created along each branch of the database glycan are checked in the experimental spectrum and a score assigned based on the appearance of contiguous fragment peaks. As previously mentioned, the observation of contiguous fragments increases the probability that the fragments are valid and that the structure represented by the theoretical spectrum is the correct match. Checking for contiguous glycan fragments is also important in the analysis of low-quality spectra in which the glycan fragment peaks may not be very intense. In this case, intensity alone may be inadequate to indicate the likelihood that a given peak represents a valid fragment. By combining information about the peak intensity with the appearance of contiguous peaks, there is further evidence of the validity of a peak. For each contiguous peak observed, a constant factor of β is added to increase the scores of well-matching glycans.

Branch Score Each branch of the glycan structure is scored separately in order to verify the glycan substructure. The score for each branch consists of three main aspects:

- The sum of all the intensities of the matched peaks. The intensities of all peaks in the spectrum which lie in a window of 1 dalton around the theoretical peak and are found to be significant (at least 3 times the level of background noise - see section 3.3.1), are summed and added to the final score.

- The number of contiguous peaks along any one branch is also incorporated into the score. A factor of qβ where q is the number of contiguous peaks observed and β a constant factor are added to the overall score.

- The ratio of the number of matched peaks to the number of peaks expected by the fragmentation of the branch. By incorporating this information, branches that are identical in composition but which are longer than the correct structure are penalized.

In formal terms, the branch score can be described as follows:

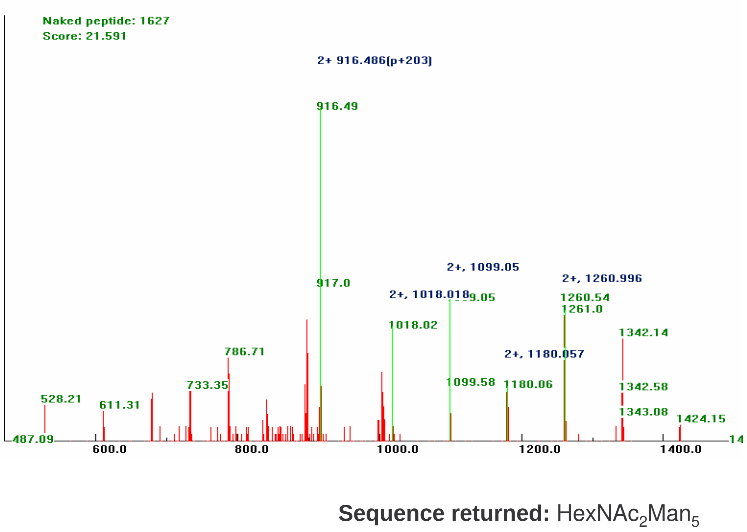
$$B = \sum_{i=0}^{i=m} intensity(i) + q\beta + (m\ peaks/number\ of\ expected\ peaks) \quad (4.1)$$

where m is the number of matched peaks and q is the number of contiguous peaks found.

The overall score for the match of the entire glycan to the experimental spectrum is taken as being the sum of all branch scores.

4.9.3 Glycan Analysis Viewer

To visualize the top glycan matches from GlycoSuite DB to experimental glycopeptide spectra, a viewer was implemented in C++. The Glycan Analysis Viewer returns a list of the top 20 matches to the experimental spectrum and indicates which peaks, and their charges, were identified in the structure. An example of the viewer is shown in figure 4.10. The viewer facilitates scientific validation of the structures returned by the Glycan Analysis Module by indicating which peaks correlate to the suggested structures.



Sequence returned: HexNAc$_2$Man$_5$

Figure 4.10: Example output from the Glycan Analysis Viewer. A high-mannose type glycopeptide was matched to a glycan with structure HexNAc$_2$Man$_5$ with the naked peptide at m/z 916.49. Matched peaks are indicated with boxes. This structure assignment was validated as being correct.

Chapter 5

Results

In this section, we discuss the performance of the Glycopeptide Classifier and the Glycan Structure Analysis Module for the tasks of glycopeptide classification from ESI-MS/MS data and N-glycan structure elucidation respectively.

5.1 Implementation

An implementation of the Glycopeptide Classifier and Glycan Analysis Module described in Chapter 3 was done in C++. This implementation was built upon existing software developed at Caprion Pharmaceuticals Inc. (Montreal) designed to analyse peptide spectra.

5.2 Input Data

A training set of 94648 spectra containing high quality glycopeptide spectra was used to develop the models described in Chapters 3 and 4. The samples, which consisted of normal and tumor tissue from patients afflicted with colon cancer, were obtained from Caprion Pharmaceuticals Inc.

Plasma membrane enriched extracts were obtained by immunoaffinity selection, and the protein extracts were separated by gel electrophoresis. Excised bands were digested by trypsin and analyzed by nano LC-MS/MS at a flow rate of 400 nL/min on a Micromass Q-TOF Ultima. The eluting peptides were ionized by electrospray and the peptide ions were automatically selected and fragmented in a data dependent acquisition mode. The MS/MS spectra were subsequently subject to data base searching for protein identification with Mascot (Matrix Science). The tissue samples were run on several machines varying in sensitivities ensuring that a variety of spectra are produced in terms of quality. The data set was examined manually for glycopeptides and a total of 188 glycopeptides were found in the sample.

5.3 Glycopeptide Classification Score Distribution

When the data data set described in section 5.2 was run through the Glycopeptide Classifier, the resulting scores were shown to be distributed as illustrated in figure 5.1a. The majority of spectra were shown to have low scores between 0 and 0.5 (fig 5.1). Although figure 5.1a suggests that no scores above 1.1 were obtained, closer examination of the scores above 1 shows a distribution of scores as seen in figure 5.1b. All glycopeptides in the sample were found to have glycopeptide classifications scores 0.8 or greater. While verifying the spectra manually, there were several spectra that were ambiguous or 'greyzone' glycopeptide spectra which could not be positively or negatively classified. These greyzone spectra were found to range from scores of 0.7-1.2 and were classified as false positives. In any one ESI-MS/MS sample, the number of greyzone glycopeptides can vary depending on the

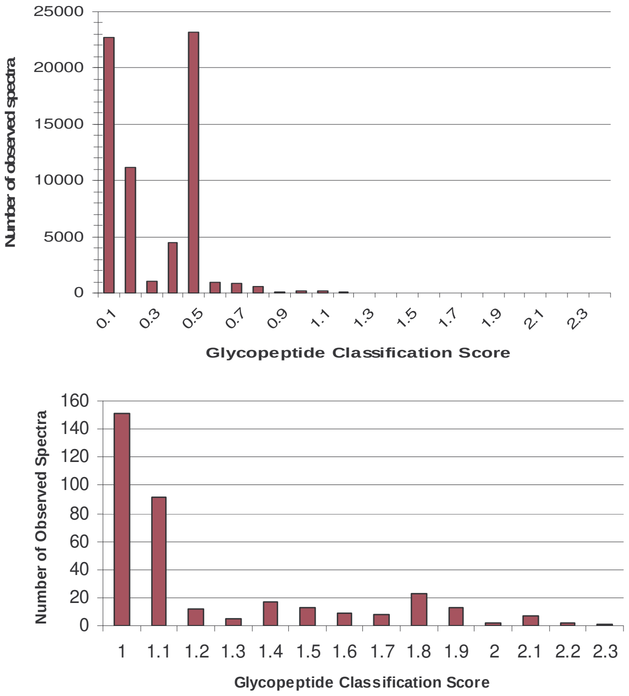


Figure 5.1: The distribution of glycopeptide classification scores for the test set are illustrated in A. Figure B illustrates the distribution of scores above 1 which do not appear in A since the number of scores at 0 greatly outnumber those above 1.

quality of the data and the parameters utilized in MS/MS (see section 3.2.3).

Based on this test set, an optimal glycopeptide decision score was established. In the next section, we will discuss the selection of the decision score as discussed in Chapter 3.

5.4 Glycopeptide Decision Score

Figure 5.1 illustrates that valid glycopeptides have a range of scores due to variations of the glycopeptide spectra for reasons discussed in section

3.2.3. To classify a spectrum as belonging to a glycopeptide, it is necessary to establish a decision score D such that if S≤D, the spectrum is not a glycopeptide and if S>D the spectrum is a glycopeptide.

There exist several methodologies for determining an accurate decision score. As a preliminary approach, we selected the decision score based on the optimal ratio of false negative to false positive glycopeptides of the data (i.e. low false negative and low false positives). The scores returned by the glycopeptide classifier were arranged into bins of 0.1 intervals, and a profile of the false negatives, false positives and true positives calculated for each bin. Figure 5.2 is a ROC plot representation of the accuracy of the classifier for classification thresholds in the range of 0.7 to 1.4. Analysis was limited to this range as below a threshold of 0.8, there were no more false negatives detected and above a threshold of 1.2 there were no more false positives detected 5.2.

Figure 5.2 illustrates that as the bins score increases, there is an increase in the number of false negatives, indicated as a percentage at each bin label. The opposite trend is observed for the number of false positives along the x-axis. The trends of figure 5.2 illustrate that in general, spectra that receive scores below 0.8 do not represent glycopeptides as there was an increase in the number of false positives with no increase in true positives since all glycopeptides had been detected at this score. Similarly, for spectra with scores greater than 1.2, there was no increase in the number of false positives and only an increase in the number of false negatives. Hits in the 0.9-1.1 range can be classified as glycopeptides with less confidence as there is a mixture of false negatives and false positives in these bins.

The results of figure 5.2 suggest that 0.9 is an optimal glycosylation score threshold. At thresholds higher than 0.9, the percentage of false negatives increases sharply. In increasing the threshold from 0.9 to 1.0, the false negative rate increases from 3.2% to 10.6%. In addition, decreasing the decision score to 0.8 increases the false positive rate from 16% for a score of 0.9 to 28% for a score of 0.8. We would like our decision score to limit the number of false negatives without creating too high a false positive rate. Since the results of the Glycopeptide Classifier will have to be manually verified, too high a false positive rate would limit the throughput of glycopeptide analysis. As such, 0.9 is an appropriate threshold as there are a relatively small number of false negatives (3%) and the false positive rate is significantly smaller than the rate at the next lowest threshold of 0.8.

5.5 Sensitivity and Selectivity of the Glycopeptide Classifier

To assess the accuracy of the glycopeptide classifier, the results were analyzed in terms of the selectivity and sensitivity of the classifier. The sensitivity of the classifier refers to the ability of the software to identify glycopeptides which are not ideal and contain phenomena like noise and missing peaks; typical obfuscating factors in MS/MS spectra interpretation. In contrast to sensitivity, selectivity refers to the ability of the classifier to distinguish spectra which represent true positives from spectra which contain some glycopeptide spectra features but represent false positives.

An initial assessment of the selectivity and sensitivity of the classifier is obtained by examining the global false positive and false negative rates. Using a threshold of 0.9, a global false positive rate of 16% and a false negative rate of 3% was observed.

Another means of assessing the selectivity of the classifier is by compar-

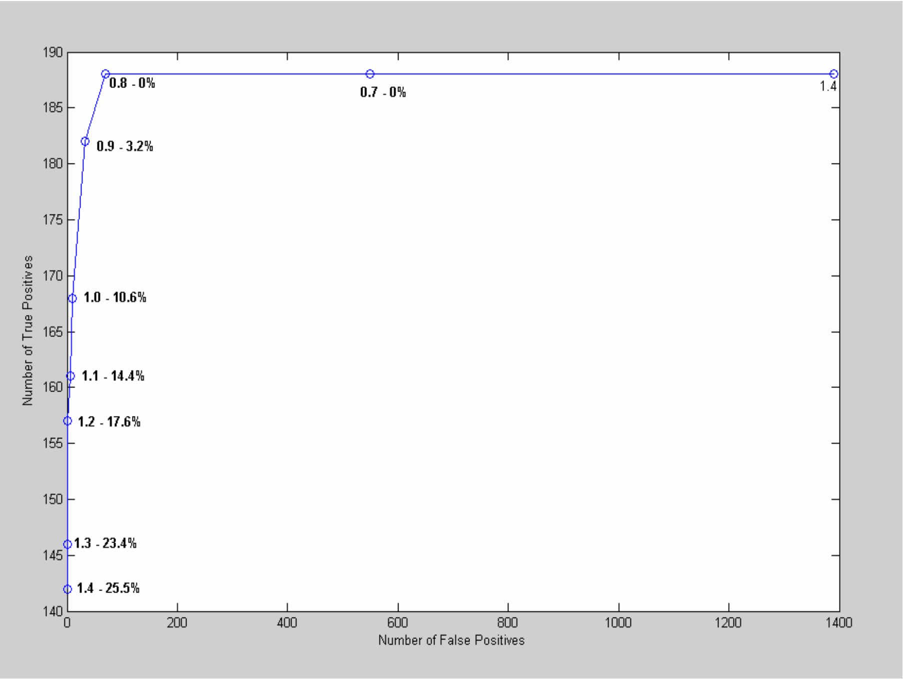


Figure 5.2: This figure illustrates a ROC plot of the Glycopeptide Classification Scores at various thresholds. The false negative rate at each decision score is displayed adjacent to the score bin label. From this figure, 0.9 was chosen as the optimal decision score as it returns the least number of false negatives while retaining a false positive score of approximately 16%.

ing the classification scores assigned to spectra of glycopeptides and non-glycopeptides. As discussed in section 3.2.2 the spectra of peptide and glycopeptide differ in terms of the criteria used to classify glycopeptides. As such, peptide spectra should receive lower glycopeptide classification scores compared to glycopeptides. In a typical high throughput proteome study, there are also a large portion of MS/MS spectra that represent neither glycopeptides nor peptides. Thus, in addition to being able to discriminate between glycopeptides and peptides, the classifier should also be able to identify random spectra as being non-glycosylated. It is likely that several spectra among the random set will contain features similar to those of glycopeptides and thus may produce higher glycopeptide classification scores than the peptide set. As such, the ability of the classifier to correctly classify the random spectra as being non-glycosylated is a measure of selectivity.

MS/MS data sets were generated to test the accuracy of the glycopeptide classifier for peptide, glycopeptide and random spectra. The peptide data were generated by obtaining MS/MS spectra which received a minimum peptide matching score of 35 from Mascot(Matrix Science) MS/MS ion searching techniques, indicating high quality peptide spectra. The glycopeptide data set consisted of all the glycopeptide spectra from the test set described in section 5.2, and the random set consisted of MS/MS spectra that were not identified as glycopeptides and which were also unassigned by Mascot MS/MS ion searching software and thus not unambiguously peptidic.

When run with on the glycopeptide classifier, the glycopeptide score distribution was shown to vary between data sets (Fig.5.3a). The glycopeptides were shown to have scores distributed between 0.9 and 2.4 with the mean glycopeptide classification score at 1.57. These results for the glycopeptide data set were shown to be significantly higher (p value < 0.01) [1] than those of the validated peptides which demonstrated a mean glycopeptide classification score of 0.26 (Fig.5.3a). No overlap between these two distributions was observed further confirming the ability of the classifier to separate glycopeptide from peptide spectra. In between the peptide and glycopeptide distributions were the scores of the random peptide sample, which demonstrated slightly higher glycopeptide classification scores than the peptide set (Fig.5.3a). Again, when the distributions of the two groups were compared, they were found to be significantly different (p<0.01). In addition, there was no significant overlap in scores between the random and glycopeptide groups although 3.4% of the random spectra received high glycopeptide scores in the range of 0.7-0.8. Upon investigation of these spectra, they were found to randomly contain glycopeptide features such as significant peaks and/or differential peak density distribution. This resolution of the random spectra from the glycopeptide spectra indicates that the glycopeptide classifier was sensitive enough to identify some noisy glycopeptide spectra (scores $<$ 1.0) and also selective enough to separate random spectra which arbitrarily contain some of the features of the glycopeptide model.

Peptide Coverage Scores To further test the results illustrated in figure5.3a, the glycopeptide classifier was also evaluated by examining peptide coverage scores. Peptide Coverage Score is a measure of the 'peptidic' quality of a spectrum and indicates the percentage of a spectrum that is spanned by amino acids and thus likely to represent a peptide spectrum. Peptide coverage scores greater than 100 generally represent peptide spectra. It is

[1] Normality plots were generated for all three groups of data. Since the peptide and random groups displayed non-normal behaviour, the Wilcoxon ranksum test (unpaired) was applied to compare the glycopeptide classification scores of the three groups.

expected that spectra that receive high glycopeptide scores should receive low peptide coverage scores and vice versa.

As in the examination of the glycopeptide classification score, peptide coverage was calculated for 3 sets of data: those displaying glycopeptide classification scores greater than 0.9, 0-0.3 and 0.4-0.9 which roughly represent glycopeptide, peptide and random spectra respectively. The peptide coverage scores for each of these data are shown in figure 5.3b.

Peptide coverage scores were shown to have an opposite trend to the glycosylation scores. The highest scores were assigned to the low glycopeptide classification score set (mean 94.5) and the lowest scores for the high glycopeptide classification score set (mean 19.2) (Fig. 5.3b). As was observed in the glycosylation score distribution, there was no overlap between the distributions of the low glycopeptide score group and the high glycopeptide score group and they were found to be independent groups ($p<0.01$).

The scores for the mid-range glycopeptide score group (mean coverage score of 56.8) were found to straddle the low and high-range glycopeptide score groups. In addition, there was significant overlap of the mid-range group with both the other groups (fig. 5.3b). This overlap between the mid-range and low-glycopeptide classification scores shows that peptide coverage score are not adequate in separating these two groups. Peptide coverage analysis reveals that the glycopeptide model is accurate and that high glycopeptide scores are selective and truly separate glycopeptide from non-glycopeptide spectra.

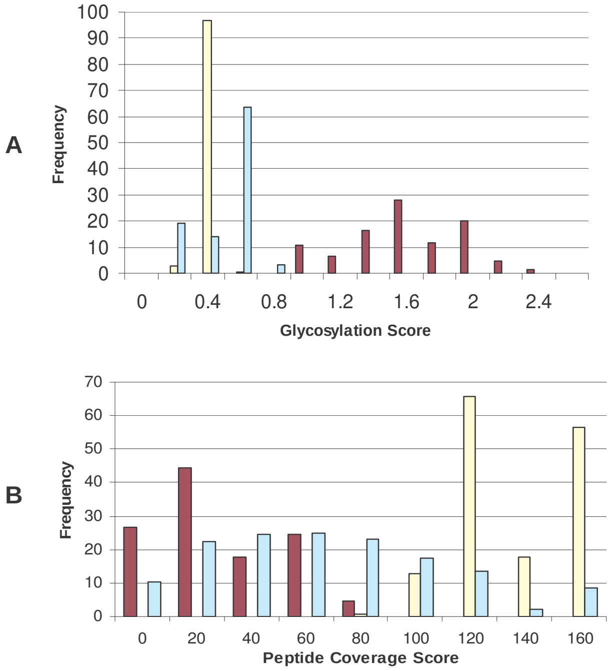


Figure 5.3: This figure illustrates the distribution of scores of three data sets: glycopeptide (dark bars), peptide (white bars), and random spectra (grey bars). A illustrates the glycopeptide classification scores for the three data sets and B illustrates the peptide coverage scores for the same data.

5.5.1 Performance Evaluation of Glycopeptide Classifier: Summary

In general, the glycopeptide classifier was found to be very effective at correctly identifying glycopeptide spectra. Out of 94648 spectra examined, the classifier was able to identify 97% (at threshold 0.9) of the true positives in the sample which equal roughly 0.2% of all spectra in the sample. At this threshold, a false positive rate of 16% was found, which also includes all greyzone glycopeptide spectra. In terms of execution time, the glycopeptide classifier was able to process 500 spectra per minute on a 1.3 gigahertz processor.

5.6 Glycan Structure Analysis Module

Both the Full and Path models of glycan fragmentation were implemented in C++ and run on the test set of glycopeptides.

The accuracy of the Glycan Analysis module will be evaluated on both the Path and Full models of fragmentation. The following criteria will be used to assess the performance of the Glycan Analysis Module:

- Percentage of hits returned with the correct naked peptide mass
- Similarity of the top structures returned to the correct structure
- Reproducibility of result in glycoforms

5.6.1 Data Sets for Glycan Analysis Module Evaluation

As discussed in Chapter 4, the performance of the Full and Path models is expected to vary depending on the type of glycan analyzed. As such, the performance of both the Full and Path models will be examined separately for both complex and oligomannose-type glycans (hybrid-type glycans were not found among the data).

Since the elucidation of manual glycan structure evaluation is a labor-intensive task, not all structure of the 188 identified glycopeptides in the test set were analyzed. In addition, the quality of many of the glycopeptide spectra obtained was low and contained many missing peaks. From the set of acquired glycopeptides, only those spectra which were of high quality were used in the evaluation of the Glycan Analysis Module. As is the case with all MS/MS ion searching techniques, the success of the matching is dependent on the spectra containing a fairly complete ladder of fragment peaks. In the selection of the high quality glycopeptide spectra, those spectra which contained more than 75% of the partial glycan fragment peaks predicted by the Path Model were used for the evaluation of the glycan analysis module. Once obtained, these spectra were pooled manually and separated into two sets depending on whether the glycan was classified as being complex or oligomannose. The oligomannose data set consisted of 15 high quality spectra and the complex data set consisted of 12 high quality spectra.

5.6.2 Accuracy of Naked Peptide Identification

The accuracy of the program to correctly identify the starting point of the glycan in the glycopeptide spectrum, the peak representing the naked peptide, was assessed by observing the percentage of correct naked peptide masses identified within a margin of one monosaccharide mass. In addition, the correct charge of the naked peptide had to be correctly identified. In general, it was found that for high quality spectra, both the Full and Path models performed equally well in determining the correct naked peptide and that the results were not contingent on the type of glycan analyzed. It was found however that for the set consisting of lower quality spectra, that the accuracy of naked peptide detection was much lower.

Specifically, the oligomannose spectra set produced 12/15 of the correct naked peptides and in the complex data set, 10/12 of the naked peptides were correctly identified. In addition, for identical glycopeptides analyzed on different machines or for glycoforms, the same naked peptide was returned. For example, glycopeptides 948.80 and 1002.76 of the oligomannose data set represent the same glycan with the higher mass glycopeptide containing one extra Hexose residue[2]. When the path model was executed on these glycoforms, peak 916.5 was identified as the correct naked peptide in both cases.

Of the incorrectly assigned naked peptides, 75% were the result of a false charge assignments for the naked peptides in the oligomannose data sets, and 100% in the complex data set. If the isotopic distributions were not well resolved, there was some ambiguity regarding the peak charge. As a result of an incorrect charge assignment to the naked peptide, all subsequent peaks were incorrectly assigned as well. An example of an ambiguous peak charge is illustrated in figure 5.4 in which the separation between the peaks fall between values of 0.33 indicating a triply charged series and 0.5, a doubly charged series. This result is likely due to the collision of 2 different isotopic distributions at the same m/z value. It is suspected that an improvement to peak charge detection, by incorporating some information about the shape of the distribution and the relative intensities of the peaks in the distribution

[2] Both the precursor 948.8 and 1002.8 are triply charged precursors. As such, a difference of 162 is produced from the calculation:(1002.76-948.80)*3-3

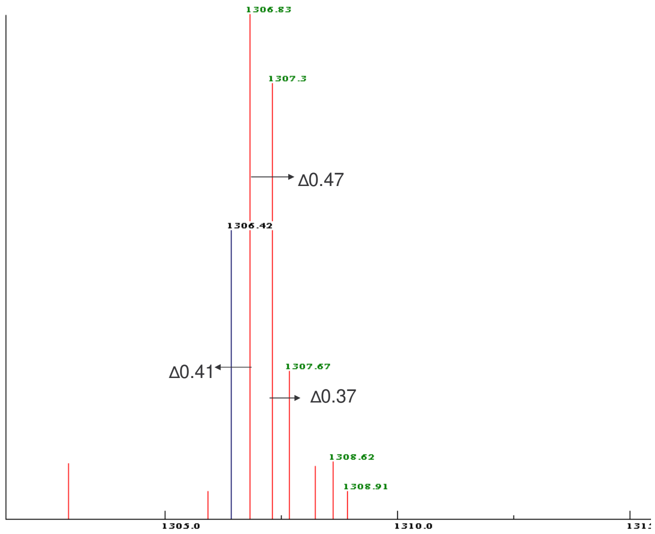


Figure 5.4: Example of an ambiguous charge assignment. In this doubly charged distribution, there is a discrepancy between the observed peak mass differences shown in the figure and the expected value of 0.5 typical of +2 isotopic distributions.

as done by Ethier et al. [11] could significantly improve the performance of the tool in this regard.

During the analysis of the naked peptide detection accuracy, it was also noted that the naked peptide peak with the additional GlcNAc residue was not always among the most intense among the peaks in the high m/z range as is typical of N-linked glyocpeptide spectra. If the actual naked peptide-GlcNAc peak is not found amongst the set of the most intense peaks, it may not be tested as a potential starting point. Figure 5.5 illustrates this phenomenon. In figure 5.5, the glycopeptide spectra of two identical glycopeptides analyzed in the same sample are shown. In figure 5.5A the peaks are not as intense as those in figure 5.5B in which the naked peptide peak is not represented. When the spectrum in 5.5B was analyzed with the Glycan Analysis Module, an incorrect naked peptide was selected as the glycan starting point of the structure resulting in an incorrect glycan assignment. This problem was found to be more prevalent in the analysis of low quality spectra which were not considered in the current analysis. Methods to improve naked peptide detection should be investigated in future versions of this program.

5.7 Glycan Structure Elucidation

In addition to the detection of the correct naked peptide, the performance of the glycan analysis module was also evaluated on its ability to return a correct monosaccharide composition and glycan structure. To evaluate the effectiveness of each fragmentation model in the elucidation of glycan structure, two main metrics were used. The first criteria examines the number of observed peaks (n_o) found in the spectrum versus the number of peaks predicted by the fragmentation models (n_e). For each glycopeptide in the complex and oligomannose data sets, the structure of the glycan was examined and the peaks representing the various partial fragments and their charges identified. These observed peaks were matched against those correctly identified (in terms of m/z and charge) by the Glycan Analysis module. This ratio of n_o/n_e provides an assessment of the accuracy of the fragmentation models. The other main metric was the ratio of the number of matched peaks (n_m) to the number of observed peaks (n_o). The ratio of n_m/ n_o provides

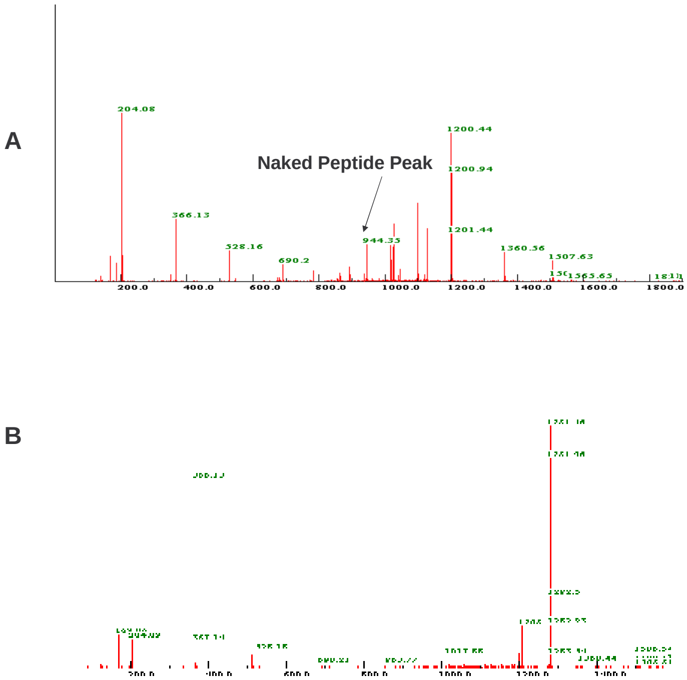


Figure 5.5: The spectra illustrated in A and B represent the same glycopeptide (from Anti-Colorectal Carcinoma Heavy Chain)analyzed twice in the same sample. The peaks in the spectrum in A are more intense than in B and the naked peptide, m/z 944.3 as shown in A, is not located in the spectrum of B. As such, an incorrect structure was reported for B when analyzed with the Glycan Analysis Module.

and assessment of the overall ability of the software to correctly identify the partial fragments in the spectrum, both in terms of the fragments generated by the fragmentation model as well as the performance of the tool in peak detection and isotope identification. The other main criteria used to evaluate the match was a qualitative assessment of the similarity of the structure of the top hits to the structure of the glycan represented in the spectrum.

In the following sections, we will examine the accuracy of the Glycan Analysis Module according to the criteria described above. The results will be examined separately in the complex and oligomannose data sets, and both the behaviour of the Path and Full Models compared in each data set.

5.7.1 Complex N-Glycan Structures

Fragmentation Models To examine the accuracy and completeness of the set of theoretical fragments produced by the Full and Path fragmentation models, the number of observed glycan fragment peaks of each spectrum in the data set was compared to the number of fragments predicted by each model (n_o/n_e). The accuracy of the glycan fragmentation models as applied to complex type glycans is summerized in Table 5.7.1. In general, the ratio of observed peaks to predicted peaks in Full Model was found to be approximately 0.32 suggesting that for the majority of complex N-glycans only a small number of predicted peaks are actually produced. This surplus in theoretical fragments is partially responsible for random peak matches obtained by the Full Model (discussed further below).

Precursor m/z	Full:nm/no	Path: nm/no	Full: no/ne	Path: no/ne
1199.91	0.875	0.635	0.241	1.0
1199.92	1.833	0.833	0.379	1.57
1199.929	0.916	0.167	0.379	1.57
1199.99	0.778	0.556	0.241	1.0
1201.99	1	0.667	0.25	0.857
1301.46	0.8	0.6	0.275	1.0
1301.46	0.889	0.778	0.275	1.0
1301.51	1.4	0.9	0.379	1.375
1301.52	1.3	0.7	0.448	1.625
1302.1	3.67	2.33	0.482	1.75
1303.5	0.667	0.667	0.166	0.5
1495.96	0.75	0.35	0.3	1.07
Average:	**1.18**	**0.764**	**0.32**	**1.19**

Table 5.7.1 Results for complex data. nm = number of matched glycan peaks, no = number of observed glycan peaks, ne = number of expected peaks from the fragmentation model.

The same ratio in the Path Model was found to be 1.19 indicating that all predicted peaks are observed. Furthermore this ratio shows that in several cases there are more peaks observed than predicted. This result can be attributed to the fact that the Path Model does not take into account branch combinations which contributes to a small number of observed peaks.

Glycan Composition In table 5.7.1, the ratio of matched peaks to observed peaks (n_m/n_o) is shown for the complex glycans using both the Full and Path models. The average value of n_m/n_o in the Full Model was computed to be 1.18 and that of the Path Model was found to be 0.76, suggesting that the Full Model was capable of identifying more partial glycan fragments in the spectrum. However, since the ratio in the Full Model is greater than 1, this result also suggests that the Full Model is matching peaks that are not observed. As discussed above, the Full Model produces more fragments than observed in the spectra. This surplus of peaks increases the likelihood of incorrectly matching theoretical fragments to noise. Random peak matches were noted in 11.5% of the Full Model matches and 7% of Path Model matches.

In addition to matching false peaks, some peak matches were missed due to incorrect charge assignments. In several cases, the theoretical fragment was found in the experimental spectrum but not matched as the isotopic distribution was incorrectly resolved (see section 5.6.2).

Glycan Structure In general, although the Path Model was able to match less peaks, the structures returned by both the Full and Path models were comparable. In figure 5.6 for example, the structures returned by both the Full and Path models for precursor 1301.51 (+2), with 9 and 14 matched peaks for the Path (Fig. 5.6b) and Full models (Fig. 5.6c) respectively, are illustrated. In figure 5.6a, the correct structure is shown.

To compensate for the branch combination masses not examined by the Path Model, the top hit returned by the Path Model (Fig. 5.6b) returns a structure with two extra fucose residues. These extra fucose residues are incorrectly matched to masses representing the combinations of the partial glycan fragments at that point (residues 5 and 6 in figure 5.6b) plus the labile core fucose molecule; masses which the Path Model does not take into account. Similarly, the glycan returned by the Path Model incorporates an additional Hex molecule (residue 3 in figure 5.6b) to account for a peak

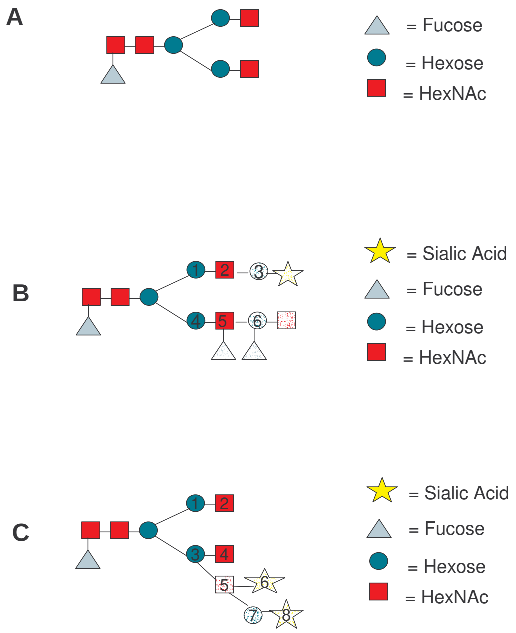


Figure 5.6: In this figure, the correct structure for the glycan of glycopeptide with m/z 1301.51 in the complex data set is illustrated in A. Figures B and C demonstrate the top hits returned by the Path and Full models respectively. Residues not matched by the program are shown as being shaded.

representing a combined fragment of residues (1+2+4). From the MS/MS spectrum, this peak may appear to be valid. However, knowledge of glycan structure and biosynthesis suggests this is not the correct structure of the glycan.

Figure 5.6c shows the top structure returned by the Full Model of fragmentation. This hit contained all monosaccharides of 5.6a and some extraneous peaks, residues 5,6,7,8. The scoring scheme did not sufficiently penalize extra monosaccharides along this branch, and the incorrect structure produced the same score as did figure 5.6a. Since the Full Model took into consideration branch mass combinations, a more accurate structure was returned compared to the Path Model.

Although the above example shows that the Full Model can return more accurate structures, the structure returned by the Path Model in figure 5.6b is still a sufficient match since residues 1,2,4,5 were identified.

Effect of Glycan Size on Performance One possible explanation for the comparable performance of the two fragmentation models for complex N-glycans could be that the glycans tested in the data set are not large enough to observe a significant difference in performance. As seen in table 5.7.1, there is a discrepancy in the number of expected peaks from the Full Model and the Path Model. For example, for precursor 1301.52, there are 14 predicted peaks for the Path Model versus 29 for the Full Model. This difference is largely attributed to the core fucose residue attached (Fig.5.6a) which contributes to half of the predicted peaks (i.e. each partial fragment with and without the fucose residue combined). Without these fucose-combined masses, the number of theoretical fragments for both models is approximately the same. However, as the size of the glycan increases and there are more branches, the number of fragments observed in the experimental spectra may become larger in the Full versus the Path Model. For larger glycans therefore, the Full Model may produce additional peaks necessary for obtaining the correct structure.

To fully observe the effect of glycan size on the accuracy of the Full and Path model would require a thorough examination of several large glycans. However, these data are currently unavailable since the glycopeptides analyzed were obtained from high throughput proteomic data in which there is an upper limit of 4000 daltons for mass. As a result, massive glycan structure were not found among the complex test set.

A independent sample of ESI-MS/MS spectra representing a well-studied glycoprotein, Bovine Fetuin (NCBI accesion: gi:27806751), was obtained and tested with both the Path and Full models. Figure 5.7a shows the MS/MS spectrum obtained for the glycopeptide and figure 5.7b shows the structure of the glycan as reported by Carr et al[3]. In the spectrum of figure 5.8, the high intensity peaks of the spectrum represent two series of peaks: one series which represents one fragmentation event per glycan (i.e. along one branch of the structure) in figure 5.7b and another series representing two fragmentation events (all peaks greater than 1360 illustrated in figure 5.8).

The number of theoretical fragments for the structure was found to be 45 and 14 for the Full Model and Path models respectively. When run with both the Full and Path models, the Full Model was shown to return a more accurate structure as seen in figure 5.9b compared to that of the Path Model in 5.9c.

In figure 5.9c, we see that only one branch of the structure is counted in the scoring. However, in addition to hits returned with this structure, a second series of hits was also returned in which a peak with the mass of

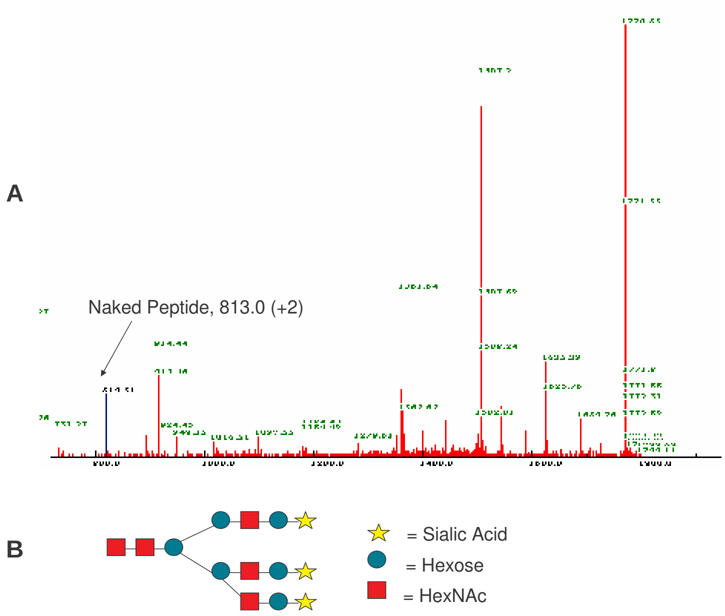


Figure 5.7: This figure illustrates the ESI-MS/MS spectrum of the Bovine Fetuin glycopeptide with m/z 1495. Figure A displays the MS/MS spectrum of the glycan moiety of the glycopeptide and B shows the real structure as elucidated by Carr et al.[3]

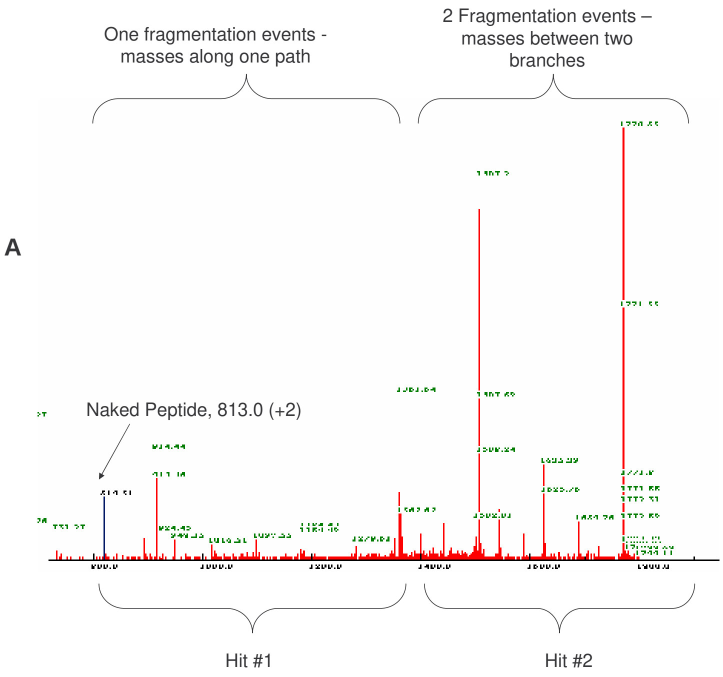


Figure 5.8: This figure illustrates the inability of the Path model to return the correct structure in the examination of large glycans. The entire structure of the glycan was elucidated by the Path Model in two hits, each representing one of the series represented in the spectrum. This problem can be attributed the fact that the Path Model does not take combined masses into account.

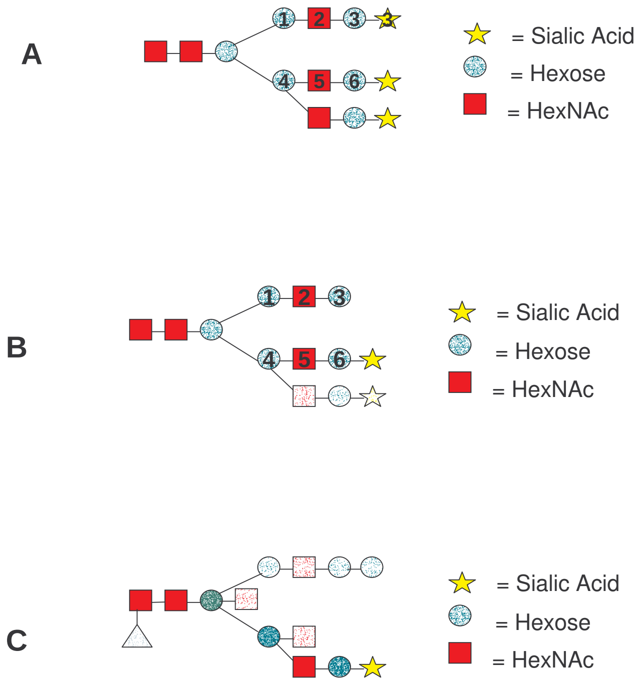


Figure 5.9: This figure shows the structures returned by the Glycan Analysis Module in the analysis of a glycopeptide from Bovine Fetuin. Figure A shows the real structure of the glycan as described by Carr et al. Figures B and C show the structures of the top hits returned by the Full and Path Glycan Fragmentation Models respectively. The shaded residues in represent residues that are incorrectly identified.

residue 4 illustrated in 5.9a served as the naked peptide. Effectively, the entire structure was elucidated in 2 hits representing series 1 in figure 5.8 and series 2 in figure 5.8.

This example illustrates that for larger glycans, the increased number of peaks produced by the Full Model may provide information necessary for the identification of the entire structure and that Path Model for large glycans may only be effective in returning substructures. Further exploration of this aspect should be carried out in the future.

Precursor m/z	Full: nm/no	Path: nm/no	Full: no/ne	Path: no/ne
876.89	2	1.14	0.467	0.636
881.75	2.14	1	0.78	1.0
917.399	0.857	0.857	0.466	0.636
948.759	0.5	1	0.67	0.857
948.77	0.714	0.714	0.67	0.857
948.786	0.429	1	0.67	0.857
948.808	1	1	0.67	0.857
948.81	2.16	1.66	0.67	0.857
1002.76	1.2	1.2	0.7	0.875
1002.78	1.67	1	0.7	0.875
1002.78	1	1	0.7	0.875
1021.27	1.125	1.125	1.11	1.25
1061.79	1.11	1	0.89	1.0
1102.307	1.11	1	1	1.125
1102.32	1.128	1.14	1	1.125
average:	**1.14**	**1.02**	**0.72**	**0.89**

Table 5.7.1 Results for oligomannose data. nm = number of matched glycan peaks, no = number of observed glycan peaks, ne = number of expected peaks from the fragmentation model.

5.7.2 Oligomannose Glycans

This section examines the results of the Glycan Analysis module implemented with both the Full and Path models of fragmentation on oligomannose type glycopeptides.

Fragmentation Models Compared to the analysis of complex glycans, the discrepancy in the ratio of observed peaks to theoretical peaks ($\mathrm{n}_o/\mathrm{n}_e$) in both models of glycan fragmentation was much smaller; ratios of 0.72 and 0.89 were observed for the Path and Full models respectively. This smaller discrepancy was expected since there is smaller variability in monosaccharide composition, and as discussed in section 5.7.1, the size of the set of peaks produced by the Full Model is smaller than in complex glycans. Thus, for oligomannose glycans, both fragmentation models performed similarly.

Monosaccharide Composition The average ratio of matched peaks to observed peaks ($\mathrm{n}_m/\mathrm{n}_o$) in oligomannose glycans was found to be 1.14 and 1.02 in the Full and Path models respectively. In all spectra of the oligomannose data, the observed peaks were correlated to partial glycan fragments in the spectra. Compared to complex glycans, there is a smaller discrepancy in ($\mathrm{n}_m/\mathrm{n}_o$) between the two models. This result is a direct consequence of the smaller number of fragments produced for oligomannose glycans.

Glycan Structure When the Path Model of fragmentation was used in the analysis of the oligomannose glycans, in all cases the correct structure was determined. The Full model of fragmentation was found to perform worse than the Path model on oligomannose sugars. Out of all the oligomannose spectra, 46% of spectra were assigned oligomannose structures among the top 5 hits returned by the Glycan Analysis Module. Although in the majority of oligomannose glycans suitable structures were returned, in 20% of cases complex glycans were returned instead of oligomannose structures. It is important to note however, that although an incorrect structure was returned, many of the observed peak assignments were correct. The difference in the performance can be partially attributed to the fact that the large number of peaks produced by the Full Model were matched to noise. In figure 5.10a the spectrum for oligomannose glycan with m/z 876.88 and composition $\text{HexNAc}_2\text{Hex}_9$ is illustrated. In the mid-high m/z range of the spectrum there is some area of noise as shown. Since the Full Model of fragmentation produces a large set of theoretical fragments, many peaks are matched to noise and as a result, one of the top hits returned by the model is a complex glycan (Fig.5.10b). To avoid this type of random matching, a more stringent measure to penalize unmatched peaks should be adopted.

5.8 Performance on Low Quality Spectra

When the Glycan Analysis Module was tested on spectra of lower quality in which only 50-70% of the partial fragment peaks are present, there was a marked degradation in the performance in both the Full and Path models of fragmentation. Figure 5.11 shows the spectra of two glycoforms from the same sample. Figure 5.11a is a high quality spectrum whereas the peaks of the spectrum in figure 5.11b contain a much smaller number of high intensity peaks. The output of the software for the spectrum in figure 5.11a produced

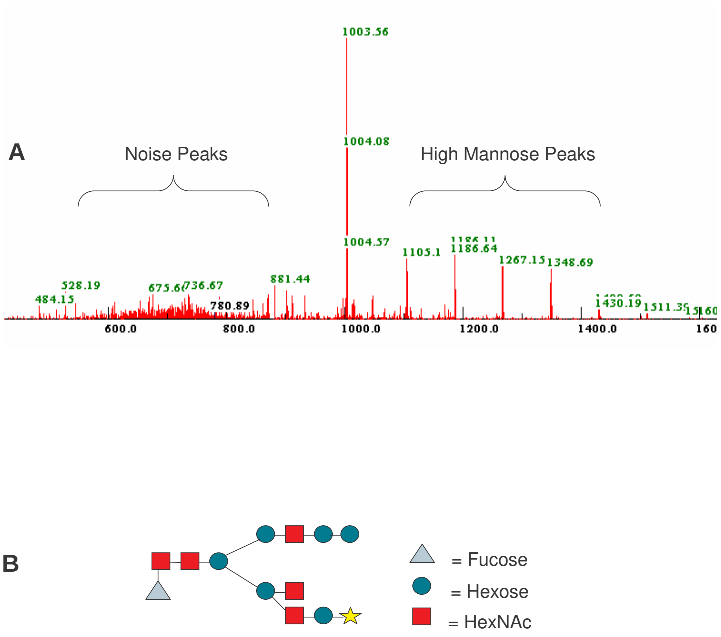


Figure 5.10: Example of matching of Full Model peaks to random noise. Figure A illustrates the spectrum for an oligomannose glycopeptide. The extraneous peaks produced by the Full Model were matched to some random peaks as shown in A. As a result of this random matching, a complex glycan as shown in B was incorrectly returned.

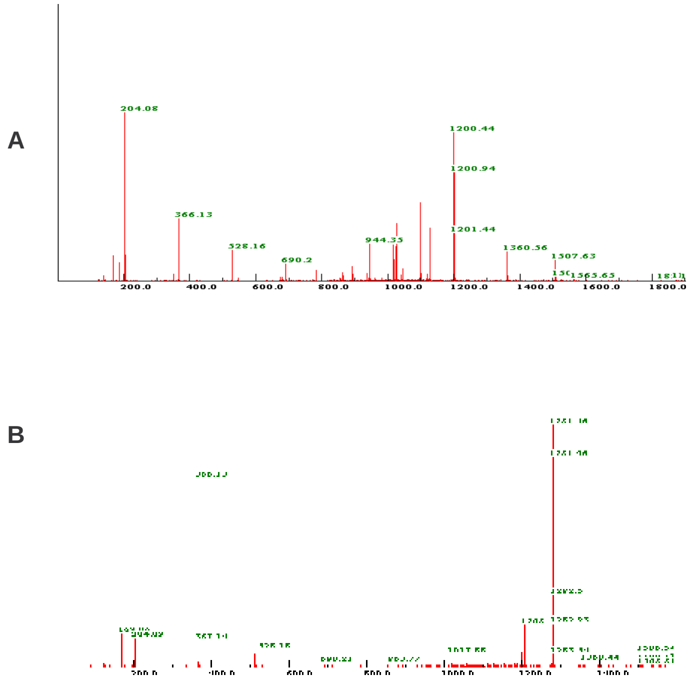


Figure 5.11: In this figure, parts A and B represent the same glycopeptide analyzed twice. However, since the quality of the spectrum in A is much higher than that of B in terms of number of partial fragment peak obtained, the structures returned for A were more accurate compared to B.

the correct glycan and the correct naked peptide. Since the peaks representing the naked peptide is not observed in the spectrum in figure 5.11b, the correct naked peptide and the glycan structure were not returned. A reduction in accuracy is also observed with MS/MS ion searching techniques when applied to low quality peptide spectra.

5.9 Execution Time

In terms of execution time, the Glycan Analysis Module performed similarly for the Path and Full Models. Although fewer peaks are produced by the Path Model, there was not a significant increase in the time required for matching. In addition, since the size of the glycans examined was small, the additional time required for the matching of larger glycans could not be examined. In general, an average of two spectra per minute were analyzed by the Path Model. A similar performance was noted for the Full Model.

5.10 Application of Glycopeptide Classifier and Glycan Analysis Module in Differential Glycoprotein Expression

The software presented in this thesis was integrated into a high throughput proteomics pipeline to assist in differential glycopeptide expression studies in normal and tumor tissue of patients afflicted with colon cancer. After MS/MS spectra for the samples were acquired, they were run through both the Glycopeptide Classifier and the Glycan Analysis Module. For a glycopeptide identified by the Glycopeptide classifier at m/z 1021.16, the MS survey scans in this m/z range were analyzed in both the normal and tumor tissues of a particular patient. Analysis of the survey scans revealed that the peptide was upregulated in tumor tissue as illustrated by the large peak at m/z 1021.16 in the tumor sample (fig.5.12b) versus the smaller peak at the same m/z in the normal sample (fig.5.12c).

To match the differentially expressed glycopeptide to its parent protein, another piece of software was implemented, the Protein ID module. The input to the Protein ID module is the mass of the naked peptide. The program then attempts to match this mass to all tryptic peptides of the NCBi database containing the NXS/T sequon common to all N-linked glycoproteins. Further, the software was enhanced to detect other PTMs and combinations of PTMs such as oxidation and methylation. For the upregulated glycopeptide, the Glycan Analysis Module suggested an oligomannose glycan structure ($\mathrm{HexNAc_2Hex_9}$) and a naked peptide m/z of 915.57. Using the protein ID module, the naked peptide of the differentially expressed peptides was matched to the protein Carcinoembryonic Antigen (CEA5_HUMAN), a known glycoprotein marker for cancer.

This example illustrates the capabilities of the software developed in this thesis to facilitate differential expression and drug target discovery in glycomics.

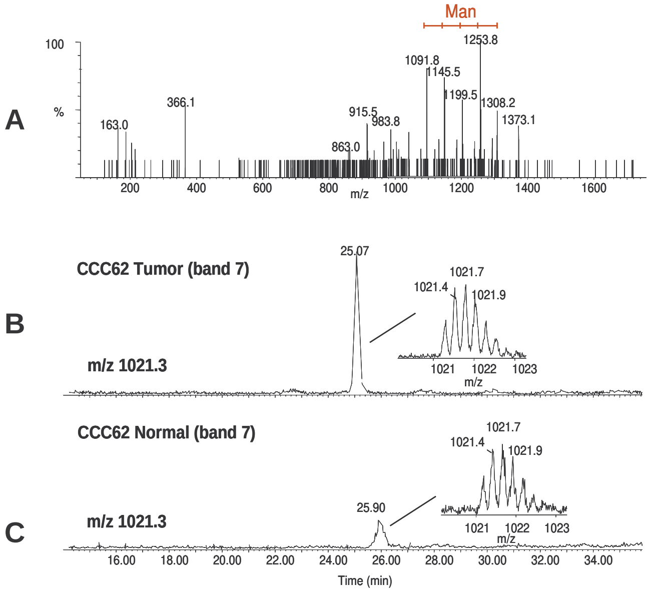


Figure 5.12: This figure illustrates the ability of the software to assist in differential glycopeptide analysis. Part A illustrates the MS/MS spectrum of a differentially expressed glycopeptide at m/z 1021.16 . Upon the examination of the survey scans of the tumor and normal tissues at this m/z range, parts B and C respectively, the intensity of the peak at 1021 was found to be much more intense in the survey scan of the tumor as opposed to that of the normal sample and thus differentially expressed. Using the Protein ID module, the glycopeptide was eventually mapped to Carcinoembryonic Antigen (CEA5_HUMAN), a known protein marker for cancer.

Chapter 6

Conclusions and Future Work

6.1 Glycopeptide Classification

The glycopeptide classification module was found to be effective in the identification of glycopeptide spectra from ESI-MS/MS data. Out of 94648 spectra examined in the test data set, 97% of all glycopeptides in the sample were detected with a false positive rate of 16%.

To improve the false negatives rate, several aspects of the classifier can be modified. Most of the missed glycopeptide spectra contained noise in the low m/z range, and the score was reduced to reflect the density of non-oxonium ion peaks in the low m/z range. It was often the case that there were valid peaks surrounding oxonium ion peaks which according to the scoring scheme developed in section 3.3.1, were to be penalized. For example, there were often two additional peaks surrounding certain oxonium ions representing two successive water losses from the ion which were regarded as extraneous peaks and penalized. Including information about these water loss peaks and other peaks to be expected in the low m/z range could reduce the false negative rate.

There are several techniques that could be applied to reduce the false positive rate. However, closer examination of the nature of the false positives showed that many of them could be classified as greyzone glycopeptide spectra that can not be positively or negatively classified manually. Thus, we can not be certain that the classes are perfectly separable and it may be necessary to introduce a third greyzone class. Without introducing a third class, it is also possible that the following techniques could increase the accuracy of classification:

- **Using more precise methods for determining the decision score**. The decision score for glycopeptide classification was determined using a straightforward approach of examining the false positive to false negative ratios at every score bin and choosing the score of the bin with the optimal ratio as the decision score. In most classification problems including that of glycopeptide classification, the classes are not perfectly separable, and at a given score S members of more than one class can exist with a given probability. To maximize the separation between classes at a specific score, we can develop a discriminant function f(x;θ) to maximize a specific measure of separation between the classes. There are several known methods to derive this function, such as Fisher's Linear discriminant analysis method or perceptrons, that may enable a more accurate classification.

  To improve the separation between classes, it is also possible to use Support Vector Machines (SVMs) which transform linear decision surfaces to more complex surfaces by extending the measurement space[13]. It is possible that in a non-linear decision space more accurate separation of the original data set is achieved. However, given that the classes may not be perfectly separable, it is uncertain whether or not SVMs would improve the classification accuracy of the Glycopeptide Classifier.

- **Incorporation of Probabilistic Information about Glycopeptide Spectra Features** For each feature of the glycopeptide model, constant weights were assigned to reward the appearance of several features. In the scoring of the oxonium ions for example, constants α and β were added to reward the appearance of a ladder of oxonium ions. The values for these constants were decided in an ad hoc manner based on the suggestions of a biologist. However, if the probability of the specific ions appearing can be assessed by surveying a large number of glycopeptide spectra, more precise weights can be assigned which may increase the performance of the Glycopeptide Classifier. During the development of the glycopeptide classifier however, we were limited to a small number of glycopeptide spectra and thus could not implement probabilistic techniques.

6.2 Glycan Analysis Module

The performance of the glycan analysis module produced mixed results depending on the type of glycan analyzed and the fragmentation model. In general, it was found that the Path Model performed marginally better than the Full Model in the analysis of oligomannose glycans. The opposite trend was observed in the analysis of complex glycans. In addition, it is postulated that the Full Model is more effective in the analysis of larger glycans. Since larger glycans produce spectra with more peaks representing mass combinations, the additional masses combinations examined by the Full Model will most likely produce more accurate results.

In terms of overall performance, neither model of fragmentation was shown to be superior to the other. To improve the performance of the glycan analysis module, a more accurate model of glycan fragmentation should be developed. The section below discusses several strategies for creating a more realistic model of glycan fragmentation.

6.2.1 Improved Model of Glycan Fragmentation

The following factors can be examined to produce a more accurate model of theoretical glycan fragmentation:

- **Incorporation of additional rules in glycan fragmentation**. The models of fragmentation developed did not include known scientific rules about glycan fragmentation such as the lability of certain residues. As mentioned in section 4.6, there are some unlikely fragmentation products that are still included in the set of theoretical fragments and which lead to random peak matches. By including known rules about fragmentation, a more valid set of peaks can be produced, increasing the likelihood of a correct match.

  The incorporation of information about the intensity of each of the glycan fragments may also help produce a more accurate model of glycan fragmentation. In general, more intense glycan fragment peaks represent common fragmentation products. For example, highly intense peaks often occur at branch points as the specific product can be derived by fragmentation along either of the branches attached at that point. They can also occur at points to which highly labile saccharides are attached. For example, masses produced by the loss of a sialic acid residue are often highly intense. Thus, if the correlation of the theoretical and experimental spectra includes information regarding intensity, it is possible that the accuracy of the hits returned may increase.

- **Determining the optimal number of fragmentation events**. The Path Model of glycan fragmentation, which produces fragments based on one fragmentation event per glycan, was proposed as an alternative to examining all possible simultaneous fragmentations across branches. Analysis of the ratios of the number of observed peaks to the number of predicted peaks by each fragmentation model however revealed that the real number of observed fragmentation seemed to lie somewhere between the Full and Path models. Although the true number of simultaneous branch fragmentation will vary according to the glycan and experimental factors, we can try to determine whether or not there is an optimal number of simultaneous fragmentation events. For this study, we can simulate the theoretical spectra that would result from 1 to m simultaneous branch fragmentations where m is the number of branches in the glycan, and compare them with observed peaks. Techniques used in Monte Carlo simulations can be adapted for this study.

6.2.2 Future Testing

The Glycan Analysis Module was tested mainly on N-linked glycans obtained from experimental data. As a result, only a small set of data was obtained with little variability in the types of glycans analyzed. In the future should the data become available, tests should be conducted on a more comprehensive data set which should include:

- Larger Glycans. See section 5.7.1.
- O-linked glycans. The approaches of both the Glycopeptide Classifier and the Glycan Analysis Module are applicable to the analysis of O-linked glycans although these spectra display slight differences from N-linked glycans. Since O-linked glycans also play a significant role in diseases such as cancer, for a complete glycome study it is important that the software is applicable to these glycans as well.
- Greater variation in the types of glycans. Glycans with different monosaccharides and structures should be tested in the future since these factors can affect glycan fragmentation patterns.

Bibliography

[1] Bafna, V., and Edwards, N. Scope: A probabilitistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics 1*, 1 (2001), 1–9.

[2] Bakthiar, R., and Tse, F. Biological mass spectrometry: a primer. *Mutagenesis 15*, 5 (2000), 415–430.

[3] Carr, S., Huddleston, M., and Bean, M. Selective identification and differentiation of *N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry.* Protein Science*, 2 (1993), 183–196.*

[4] Chen, T., Kao, M, T., Tepel, M., Rush, J., and Church, G. *A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. In* Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms *(2000), pp. 389–398.*

[5] Cooper, C., Gasteiger, E., and Packer, N. *Glycomod - a software tool for determining glycosylation compositions from mass spectrometric data.* Proteomics*, 1 (2001), 340–349.*

[6] Cooper, C., Harrison, M., Wilkins, M., and Packer, N. *Glycosuitedb: a new curated relational database of glycoprotein glycan structures and their biological sources.* Nucleic Acids Research 29*, 1 (2001), 332–335.*

[7] Dainese, P., and James, P. Proteome Research: New Frontiers in Functional Genomics*. Springer-Verlag, Berlin, New York, 1997, ch. Protein Identification by Peptide-Mass Fingerprinting, pp. 103–123.*

[8] Dell, A., and Morris, H. *Glycoprotein structure determination by mass spectrometry.* Science 291 *(2001).*

[9] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. Biological Sequence Analysis*. Cambridge University Press, Cambrige, UK, 1998, ch. Pairwise Alignment, pp. 12–45.*

[10] Dwek, R. *Glycobiology: Toward understanding the function of sugars.* Chemical Review 96 *(1996).*

[11] Ethier, M., Saba, J., Ens, W., Standing, K., and Perreault, H. *Automated structural assignment of derivatized complex n-linked oligosaccharides from tandem mass spectra.* Rapid Communications in Mass Spectrometry 16 *(2002), 1743–1754.*

[12] Gaucher, S., Morrow, J., and Leary, J. *Stat: A saccharide topology analysis tool used in combination with tandem mass spectrometry.* Analytical Chemistry 72 *(2000).*

[13] Hand, D., Mannila, H., and Smyth, P. Principles of Data Mining*. MIT Press, Cambrige, Massachusetts, 2001.*

[14] James, P. Proteome Research : Mass Spectrometry. *Springer-Verlag, Berlin, New York, 2001, ch. Mass Spectrometry and the Proteom, pp. 1–9.*

[15] Johnson, R. Proteome Research : Mass Spectrometry. *Springer-Verlag, Berlin, New York, 2001, ch. Automated Interpretation of Peptide Tandem Mass Spectra and Homology Searching, pp. 165–185.*

[16] Jonsson, A. *Mass spectrometry for protein and peptide characterization.* Cell. Mol. Life Sci. 58 *(2001), 868–884.*

[17] Jun, H., Arata, Y., and Kasai, K. *Glycome project : Concept, strategy and preliminary application to caenorhabditis elegans.* Proteomicsy, *1 (2001), 295–303.*

[18] Kennedy, M. *2nd annual emsl workshop on structural genomics. [online]http://www.emsl.pnl.gov:2080/docs/msd/workshop2000/general.htm, 2000.*

[19] Khoo, K. *Towards a rational glycomic approach in the age of functional genomics. [online]http://www.sinica.edu.tw/ kkhoo/research.html, 2002.*

[20] Lohmann, K., and von der Lieth, C. *Glyco-fragmnent: A web tool to support the interpretation of mass spectra of complex carbohydrates.* Proteomics*, 3 (2003), 2028–2035.*

[21] Ma, B., Zhang, K., Lajoie, G., Doherty-Kirby, C., Hendrie, C., Liang, C., and Li, M. *Peaks:powerful software for peptide de novo sequencing by tandem mass spectrometry.* Rapid Communication in Mass Spectrometry 17*, 20 (2003), 2337–2342.*

[22] Mizuno, Y., and Sasagawa, T. *An automated interpretation of maldi/tof postsource decay spectra of oligossacharides.1.automated peak assignment.* Analytical Chemistry 71 *(1999).*

[23] M.R., W., and Gooley, A. Proteome Research: New Frontiers in Functional Genomics. *Springer-Verlag, Berlin, New York, 1997, ch. Protein Identification in Proteome Projects, pp. 35–64.*

[24] Papayannopoulos, I. *The interpretation of collision-induced dissociation tandem mass spectra of peptides.* Mass Spectrometry Review*, 1 (1995), 49–73.*

[25] Perkins, D., Pappin, D., Creasy, D., and Cottrell, J. *Probability-based protein identification by search sequence databases using mass spectrometry data.* Electrophoresis*, 20 (1999), 3551–3567.*

[26] Rudd, P., Colominas, C., R. L., Murphy, N., Hart, E., Merry, A., Heberstreit, H., and Dwek, R. Proteome Research : Mass Spectrometry. *Springer-Verlag, Berlin, New York, 2001, ch. Glycoproteomics : High-Throughput Sequencing of Oligosaccharide Modifications to Proteins., pp. 207–228.*

[27] Stein, S. *An integrated method for spectrum extraction and compound identification from gc/ms data. [online] http://chemdata.nist.gov/mass-spc/amdis/AutoGC.html, Accessed 26 April 2002 2002.*

[28] Tabb, D, L., Eng, J, K., and Yates, J. I. Proteome Research : Mass Spectrometry. *Springer-Verlag, Berlin, New York, 2001, ch. Protein Identification by SEQUEST, pp. 125–142.*

[29] THOMAS, R. *Recent developments in lc-ms-ms for the identification and measurement of nanoscale amounts of proteins and peptides.* Spectroscopy 16*, 1 (2001), 28–37.*

[30] VARKI ET AL., . Essentials of Glycobiology*. Cold Springs Harbor Laboratory Press, La Jolla, California, 1999.*

[31] WEI, X., DECKER, J., WANG, S., HUI, H., KAPPES, J., WU, X., SALAZAR-GONZALES, J., SALAZAR, M., KILBY, J., SAAG, M., KOMAROVA, N., NOWAK, M., HAHN, B., KWONG, P., AND SHAW, G. *Antibody neutralization and escape by hiv-1.* Nature 422*, 6929 (2003), 307–12.*

[32] WHITEHOUSE, C., BURCHELL, J., GSCHMEISSNER, S., BROCKHAUSEN, I., LLOYD, K., AND TAYLOR-PAPDIMITRIOU, J. *Transfected sialyltransferase that is elevated in breast cancer and localizes to the medial/trans-golgi apparatus inhibits the development of core-2-based o-glycans.* Journal of Cell Biology 137 *(1997), 1229–1241.*

[33] ZHANG, N., AEBERSOLD, R., AND SCHWIKOWSKI, B. *Probid: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data.* Proteomics 20*, 10 (2002), 1406–12.*

Glossary

Anomericity The a or b configuration of the glycosidic bond of a sugar to another sugar.

b-ion Peptides are made of repeating amino acid units as illustrated in figure 2.1. To differentiate the start of the peptide from the end, each terminus of the peptide is labeled; one terminus is called the N-terminus (amino terminus) and the other the C-terminus (carboxyl terminus). The N terminus is where protein synthesis is initiated and the C terminus is where it is terminated. When the amino acids of the peptide fragment, they produce 2 types of ions: a b-type and a y-type ion (there exist other types of ions as well such as x-ions and a-ions although they are more rare). A b-ion and a y-ion represent the peptide fragment containing the N- or C-terminus respectively. Every b- and y-ion pair is complementary and the combined masses of the b-ion and the y-ion should equal the mass of the peptide. The interpretation of peptide MS/MS spectra thus begins with the identification of a series of ions (b or y), between which there is a mass difference equal to a specific amino acid. By detecting and entire series of b-or y-ions and observing mass differences between them, the sequence of the peptide can be reconstructed.

Carbohydrate Same definition as glycan.

Collisional Induced Dissociation (CID) This is a technique whereby precursor ions are made to undergo collision with a neutral gas to produce fragments.

Complex-type glycan N-linked glycan that has varied composition and a variable number of antannae stemming from the core (see figure 1.5B). Complex type N-glycans show the largest structural variation resulting from the combination of monosaccharides and the different number of antannae.

de novo peptide sequencing The elucidation of peptide sequences from an MS/MS spectrum directly. By observing distances between peaks in an MS/MS spectrum, the order of peptide fragmentation, and thus the sequence can be deduced.

Deconvolution For making MS/MS more easy to interpret, often the multiply charged peaks are converted into their singly charged forms. This process is MS/MS spectra deconvolution.

Electrospray Ionization (ESI) ESI is an ionization technique using ion evaporation. The sample is dissolved in a mobile phase and pumped through a capillary. The sample is then floated at high potential which confers a charge to the peptide. The ionized peptides are then subject to MS. ESI ionization is commonly used in proteomics, and was the mode of ionization used to generate the data in this thesis.

Glycan General term used to refer to a di-,oligo-, or polysaccharide structure. In the context of glycoproteins, the glycans are the carbohydrate moieties attached to the peptides.

Glycosidic bond The bond between a sugar and an alcohol. Also the bond that links two sugars in disaccharides, oligosaccharides and polysaccharides.

Glycoform When a single glycosylation site contains a series of different carbohydrates (microheterogeneity), the glycoproteins with different glycans are glycoforms of each other.

Greyzone spectra Those MS/MS spectra in which the underlying sequence is ambiguous, even upon manual elucidation.

High-mannose-type glycan See definition for oligomannose-type glycan.

Hybrid-type glycan Hybrid-type N-glycans have the characteristic features of both complex-type and high-mannose type glycans as seen in figure 1.5C.

Intensity (peak intensity) Y-axis measurement on an MS/MS spectrum. For a fragment at a particular m/z, the intensity provides a measure of the relative quantity of the peptide detected.

Isotope Atoms with the same atomic number differing in mass by one and possessing nearly identical chemical properties. In the context of MS/MS spectra, the various isotopes of the fragments are often found in a series separated by 1 dalton. If the initial fragment was multiply charged with a charge m, there will be a series of peaks spaced by 1/m m/z units.

Lability Lability refers to the ability of a particular monosaccharide to dissociate during CID.

Liquid chromatography (LC) Process used to separate complex mixtures of peptides or proteins according to various factors such as hydrophobicity. LC is often used in tandem with mass spectrometry; peptides separated by LC enter the mass spectrometer in a coupled fashion.

Macroheterogeneity Macroheterogeneity refers to the phenomenon in which a protein can have one or more glycosylation sites, which may or may not be occupied by a carbohydrate.

Monosaccharide The base unit forming glycans. The most common types of monossacharides are Hexoses (Man,Glc, Gal), HexNAc (GalNAc,GluNAc) and NeuAc (sialic acid) and pentose. Multiple monosaccharide molecules can be linked together in chains, to form disaccharides, trisaccharides, and polysaccharides.

Microheterogeneity A particular glycoprotein may occur in forms that differ in the structure of one or more of its carbohydrate units, a phenomenon known microheterogeneity.

m/z The mass to charge ratio of an ion with ”z” being the exact integer multiple of the charges on the ion. Since MS/MS spectra contain charged species, the values of a fragment are reported as m/z.

Naked peptide Term used to signify the peak representing the peptide without the glycan. During low-energy CID of glycopeptides, the naked peptide peak itself is not always visible since the peptide moiety remains intact with the first HexNAc of the glycan attached. The naked peptide peak+HexNAc is typically one of the most intense peaks in a glycopeptide MS/MS spectrum.

Oligomannose-type glycan N-linked glycan in which all monosaccharides attached to the common core of $\mathrm{HexNAc_2Hex_3}$ are mannose residues (see figure 1.5A).

Oligosaccharide A saccharide of a small number of monosaccharides, either O or N linked to the next sugar.

Oxonium ion A class of salts derived from certain organic ethers or alcohols by adding a proton to the oxygen atom and thus producing a positive ion. In the context of glycopeptide ESI-MS/MS spectra, oxonium ions are the peaks at low m/z values which represent the monosaccharides or di-, tri-saccharides (e.g. peaks at m/z 204 HexNAc, 366 HexNAcHex). Oxonium ions are a key feature of glycopeptide MS/MS spectra.

Polysaccharide Polymers of more than ten monosaccharide residues linked in branched or unbranched chains.

Sugar Same definition as glycan.

Survey scan After the peptides of a particular sample are ionized and subject to the first round of mass spectrometry, the masses and amount of each peptide is recorded in a survey scan. Peptides from the survey scan are further selected for a second round of MS for sequence determination.

y-ion See definition for b-ion.