MS/MS ANALYSIS AND AUTOMATED TOOL DEVELOPMENT FOR PROTEIN POST-TRANSLATIONAL MODIFICATIONS

By

Carrie L. Woodin

Submitted to the graduate degree program in Chemistry and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson: Dr. Heather Desaire

Dr. Susan Lunte

Dr. Michael Johnson

Dr. Timothy Jackson

Dr. Emily Scott

Date Defended: April 16, 2013
The Dissertation Committee for Carrie L. Woodin certifies that this is the approved version of the following dissertation:

MS/MS ANALYSIS AND AUTOMATED TOOL DEVELOPMENT FOR PROTEIN POST-TRANSLATIONAL MODIFICATIONS

Chairperson: Dr. Heather Desaire

Date Approved: April 16, 2013
Protein post-translational modifications (PTMs) are important for a variety of reasons, as PTMs confer the final protein product and biological functionality onto a nascent protein chain. Two of the most common PTMs are glycosylation and disulfide bond formation. Both glycosylation and disulfide bond formation contribute to a range of cellular processes, including protein folding and stabilization. Mass spectrometry (MS) has shown to be an essential technique to study PTMs, especially when tandem mass spectrometry (MS/MS) experiments are performed. In the characterization of PTMs using MS/MS, different fragmentation techniques are often used. Regardless of the dissociation method that is employed, MS/MS data interpretation is a tedious and lengthy procedure. To render this analysis more efficient, the use of automated tools is necessary.

In this work, collision induced dissociation (CID) MS/MS experiments were carried out in order to create a set of fragmentation rules applicable to any N-linked glycopeptide. These rules were then used to develop an algorithm to power publicly available software that accurately determines glycopeptide compositions from MS/MS data. This program greatly reduces the time it takes researchers to manually assign the identity of an N-linked glycopeptide to an acquired CID spectrum. In addition, electron transfer dissociation (ETD) experiments were performed in order to devise a computational approach that works to determine precursor charge state directly from MS/MS data of peptides containing disulfide bonds. Lastly, alternate fragmentation patterns found to be detected in MS/MS data of glycopeptides containing labile monosaccharide residues such as sialic acid, are discussed. These patterns, along with other trends noticed after extensive analysis of N-linked glycopeptide CID spectra, were then used to propose future updates to the GPG analysis tool.
ACKNOWLEDGEMENTS

I would first like to thank everyone in the Desaire group, past and present, for making this journey possible. All of the struggles and divergent point of views were just as important as the support and intellectual discussions, in motivating me to get where I am today. Heather Desaire, I am so grateful for your leadership in the role of my advisor. If you had not been there to challenge me to be always give my best, praise me when I performed well, and support me even when I stumbled, I could not have completed my doctoral degree. Zhikai Zhu, I will forever be grateful for the extra help and guidance, including those thoughtful conversations you so enthusiastically provided. Melinda Toumi, I want to thank you for all of the stimulating and provoking discussions, even though we sometimes clashed in opinion. You were truly a friend and mentor, especially at the end when I needed it the most.

To all of the professors who instilled in me a love of science, thank you for feeding my curiosity and fueling my determination. This is particularly true of Dr. Gary Trammell, my undergraduate advisor. If not for you, I would have abandoned my goal of a double major, and subsequently my passion for chemistry, long before my journey at KU began.

To all my family and friends, thank you for your continued support over the years. My interactions with you have shaped me into the person that I am today. At critical points over the past few years, it was you who enabled me to forge ahead. I love and appreciate you all.

To my daughter Cashylon, mommy honestly had the roughest year of her life before she learned of your precious existence; you gave me the strength to find myself again when I felt I had none left, and the courage to never give up. I still remember how I could hardly wait to meet you, and hold you preciously in my arms for the first time. It was you, Cashy, that provided the single most powerful influence on resurrecting the ambition, passion, and focus I needed in
completing my degree. Though you also made finishing graduate school physically draining, I would not change your presence or timing for anything. If there is one thing that you never have to doubt, no matter what struggles we both face, it is that you will never be alone in life…ever. I promise that you will always have me, and you will always know how much you are loved!!!
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction.</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Protein post-translational modifications</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Overview of post-translational modifications</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Characterization of protein PTMs</td>
<td>1</td>
</tr>
<tr>
<td>1.1.3 Analysis of protein PTMs by mass spectrometry</td>
<td>2</td>
</tr>
<tr>
<td>1.1.3.1 Electrospray ionization mass spectrometry</td>
<td>3</td>
</tr>
<tr>
<td>1.1.3.2 Fourier transform ion cyclotron resonance mass spectrometry</td>
<td>4</td>
</tr>
<tr>
<td>1.1.4 Analysis of protein PTMs by tandem mass spectrometry</td>
<td>5</td>
</tr>
<tr>
<td>1.1.4.1 Collision induced dissociation tandem mass spectrometry</td>
<td>5</td>
</tr>
<tr>
<td>1.1.4.2 Electron transfer dissociation tandem mass spectrometry</td>
<td>7</td>
</tr>
<tr>
<td>1.2 Protein glycosylation</td>
<td>8</td>
</tr>
<tr>
<td>1.2.1 Overview of glycosylation</td>
<td>8</td>
</tr>
<tr>
<td>1.2.2 Glycosylation heterogeneity</td>
<td>8</td>
</tr>
<tr>
<td>1.2.3 Types of protein glycosylation</td>
<td>8</td>
</tr>
<tr>
<td>1.2.4 Characterization of protein glycosylation</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Glycosylation analysis by mass spectrometry</td>
<td>9</td>
</tr>
<tr>
<td>1.3.1 Automated analysis of released glycans</td>
<td>12</td>
</tr>
<tr>
<td>1.3.1.1 Characterization of high resolution MS oligosaccharide data</td>
<td>13</td>
</tr>
<tr>
<td>1.3.1.2 MS/MS approaches for glycan characterization</td>
<td>14</td>
</tr>
<tr>
<td>1.3.2 Automated analysis of glycopeptides</td>
<td>18</td>
</tr>
<tr>
<td>1.3.2.1 Experimental data requirements</td>
<td>20</td>
</tr>
<tr>
<td>1.4 Automated MS and MS(^n) analysis of glycopeptides</td>
<td>20</td>
</tr>
<tr>
<td>1.4.1 N-linked glycopeptides</td>
<td>20</td>
</tr>
<tr>
<td>1.4.1.1 N-linked glycopeptide characterization from MS data</td>
<td>21</td>
</tr>
<tr>
<td>1.4.1.2 N-linked glycopeptide characterization from MS/MS data</td>
<td>22</td>
</tr>
<tr>
<td>1.4.2 O-linked glycopeptides</td>
<td>28</td>
</tr>
<tr>
<td>1.4.2.1 Mucin-type O-linked glycosylation</td>
<td>29</td>
</tr>
<tr>
<td>1.4.2.2 O-linked glycopeptide characterization from MS data</td>
<td>29</td>
</tr>
<tr>
<td>1.4.2.3 O-linked glycopeptide characterization from MS/MS data</td>
<td>30</td>
</tr>
<tr>
<td>1.5 Protein disulfide bond formation</td>
<td>31</td>
</tr>
<tr>
<td>1.5.1 Overview of disulfide bonds</td>
<td>31</td>
</tr>
<tr>
<td>1.5.2 Types of disulfide bonding</td>
<td>31</td>
</tr>
<tr>
<td>1.5.3 Characterization of protein disulfide bonds</td>
<td>32</td>
</tr>
<tr>
<td>1.5.4 Disulfide-bonded peptide analysis by mass spectrometry</td>
<td>33</td>
</tr>
<tr>
<td>1.5.5 Disulfide-bonded peptide analysis by tandem mass spectrometry</td>
<td>33</td>
</tr>
<tr>
<td>1.5.5.1 Automated MS/MS data analysis of disulfide-bonded peptides</td>
<td>33</td>
</tr>
<tr>
<td>1.6 Concluding Remarks</td>
<td>34</td>
</tr>
</tbody>
</table>
1.7 Acknowledgements 35

1.8 Summary of subsequent chapters 36

1.9 References 38

2. **Collision induced dissociation behavior of N-linked glycopeptides.** 50

2.1 Introduction 52

2.2 Experimental 54
   2.2.1 Materials and reagents 54
   2.2.2 Preparation of RNase B, asialofetuin, and transferrin glycopeptides 54
   2.2.3 Direct injection mass spectrometry 55
   2.2.4 Liquid chromatography and mass spectrometry 55
   2.2.5 Manual data analysis 56

2.3 Results and discussion 57
   2.3.1 Collision induced dissociation MS/MS studies 57
   2.3.2 N-linked glycopeptide fragmentation rules 63
   2.3.3 Initial algorithm development 65

2.4 Concluding remarks 70

2.5 Acknowledgements 70

2.6 References 71

3. **GlycoPep Grader: A web-based utility for assigning the composition of N-linked glycopeptides.** 74

3.1 Introduction 75

3.2 Experimental 77
   3.2.1 Materials and reagents 77
   3.2.2 Production of CON-S gp140 CFI glycoprotein 77
   3.2.3 Preparation and LC-MS of CON-S gp140 CFI glycopeptides 78
   3.2.4 Development of a glycopeptide training data set 78
   3.2.5 The glycopeptide validation data set 79
   3.2.6 Software platform 79
   3.2.7 Generation and input of glycopeptide candidate compositions 80
   3.2.8 False discovery rate determination and scoring of candidate compositions 81

3.3 Results and discussion 81
   3.3.1 Novel GPG scoring algorithm 82
   3.3.2 Candidate composition scoring by GPG 83
3.3.3 GPG validation: Application to recombinant gp120 HIV envelope protein 95

3.4 Concluding remarks 105

3.5 Acknowledgements 106

3.6 References 107

4. Computational method to determine precursor charge state in ETD MS/MS data of disulfide-bonded peptides. 110

4.1 Introduction 111

4.2 Experimental 113
   4.2.1 Materials and reagents 113
   4.2.2 Protease digestion 114
   4.2.3 Mass spectrometry on an ESI-LTQ Velos 114
   4.2.4 Manual data analysis 115

4.3 Results and discussion 115
   4.3.1 Low resolution ETD MS/MS data of peptides containing disulfide bonds 116
   4.3.2 Method development and design 118
   4.3.3 Precursor charge state assignment of disulfide ETD MS/MS data 122
   4.3.4 Number of charge state assignments returned 127

4.4 Concluding remarks 130

4.5 Acknowledgements 131

4.6 References 132

5. Future direction: GlycoPep Grader updates. 135

5.1 Introduction 136

5.2 Experimental 138
   5.2.1 Materials and reagents 138
   5.2.2 CID MS/MS data of RNase B, asialofetuin, and transferrin glycopeptides 138
   5.2.3 CON-S gp140 CFI preparation and CID MS/MS data 139
   5.2.4 Manual data analysis 139

5.3 Results and discussion 140
   5.3.1 Peptide-containing glycopeptide product ions 142
   5.3.2 Sialylated glycopeptides 143
   5.3.3 Glycopeptide marker ion detection for complex/hybrid type glycans 147
   5.3.4 Other potential future updates 153
5.4 Concluding remarks 154
5.5 Acknowledgements 155
5.6 References 156

6. Conclusion. 159

6.1 Summary of dissertation content 159
6.2 References 161
Some material in this dissertation may be subject to copyright law.
CHAPTER 1

INTRODUCTION

The work described in Chapter 1 encompasses an original (first author) publication:


1.1 POST-TRANSLATIONAL MODIFICATIONS

1.1.1 Overview of Post-Translational Modifications. There are currently hundreds of known protein post-translational modifications, commonly referred to as PTMs, that have been classified among the archaea, prokaryotes, and eukaryotes. Some of the more frequent PTMs are acetylation, glycosylation, phosphorylation, and disulfide bond formation. Collectively, these modifications work to regulate protein structure and function, and various cellular processes. The development and progression of cancer and other diseases are also shown to be influenced by protein PTMs. In eukaryotic organisms, glycosylation and disulfide bond formation are two of the most prevalent modifications that a protein undergoes after translation. As it stands, it is estimated that over 50% of a eukaryote’s cells are glycosylated and contain disulfide bonds. Furthermore, both glycosylation and disulfide bond formation have been heavily implicated in the design of safe and effective protein pharmaceuticals.

1.1.2 Characterization of Protein PTMs. Many methods of instrumental and biochemical analysis have been used to study protein PTMs, including glycosylation and disulfide bond formation, with varying degrees of success. Some examples reported in the literature are: Edman degradation, crystallography, nuclear magnetic resonance (NMR), circular dichroism (CD), and mass spectrometry (MS). Currently, mass spectrometry is
the most utilized analytical route in the sequencing of proteins containing PTMs.\textsuperscript{16, 21} In contrast to other spectroscopic methods such as NMR and crystallography, MS experiments allow for the accurate detection and characterization of these modifications using only a small volume of sample.\textsuperscript{22} Furthermore, liquid chromatography mass spectrometry (LC-MS) permits the interrogation of PTMs when heterogeneous protein mixtures are considered.\textsuperscript{22, 23, 24} To this end, LC-MS analysis of proteins and peptide PTMs in complex biological matrices is now routinely performed.\textsuperscript{22, 23}

1.1.3 Analysis of Protein PTMs by Mass Spectrometry. In the identification of protein PTMs, MS experiments have shown to be useful in mapping their location, and correlating that information with biological functionality.\textsuperscript{22, 24} This has been accomplished using various types of mass spectrometers.\textsuperscript{16, 18, 21, 22, 23, 25} Regardless of the type of mass measurements performed, there are three main components to an MS instrument.\textsuperscript{22}

1) An ion source to generate ions.

2) A mass analyzer to separate ions, based on their \( m/z \).

3) A detector to quantify each ion’s abundance.

For the study of biomolecules, electrospray ionization (ESI) and matrix-assisted desorption ionization (MALDI) are the two most common methods used to generate ions.\textsuperscript{16, 21, 22, 26} A variety of mass analyzers may be coupled to these ionization sources.\textsuperscript{16, 21, 22} ESI is routinely coupled to LC-MS systems with both low resolution, and high resolution, analyzers.\textsuperscript{22, 23} Depending on the type of mass analyzer, the method of detection for the ions varies, although in most instruments ions are detected when they make physically contact with the detector, which often occurs in electron multipliers.\textsuperscript{21, 25}

In order to distinguish isobaric peptide compositions using mass spectrometry, including
those containing post-translational modifications (e.g. glycosylated or disulfide-bonded peptides), tandem mass spectrometry (MS/MS) experiments are usually required. The fragmentation profiles obtained within tandem mass spectra allow a precursor’s identity to be determined in cases where a number of potential compositions have the same nominal mass; that is, when their in silico mass calculations differ by less than the instrument’s accepted error range.27, 28

To increase the efficiency at extracting relevant PTM information, researchers have developed an arsenal of analysis tools and computer software programs to automate both MS and MS/MS data interpretation.29, 30, 31 Figure 1 provides an illustration of this analytical process.

**Figure 1.** In the characterization of glycosylation PTMS, glycopeptide analysis provides location information to identify where the modification resides along the amino acid sequence. After the appropriate MS scans are performed, the use of computer software and automated analysis tools substantially increases the amount of data that can be processed for a given amount of time, as compared to data analyzed manually.29

1.1.3.1 **Electrospray Ionization Mass Spectrometry.** The application of ESI in characterization of biomolecules by MS has greatly advanced the field of proteomics. The wide applicability of electrospray ionization mass spectrometry (ESI-MS) to large molecules has enabled the amino acid sequencing of proteins, including those with PTMs.23, 26, 32, 33, 34, 35, 36 In ESI-MS, a sample is introduced into the source by way of a stainless steel capillary needle (ESI needle), which is heated to a temperature between approximately 200 and 500 °K.34, 35, 37, 38
Droplets containing the charged analyte are formed during the ESI process, which can be performed in positive or negative mode.\textsuperscript{16, 21} For proteins and peptides, this analysis is typically performed in positive ion mode, due to the ability of the amino groups along the primary sequence to undergo protonation.\textsuperscript{24}

In positive mode ESI-MS, charge separation of the positively and negatively charged species (these contain the analyte) occurs as the potential applied across the heated capillary attracts the positively charged species.\textsuperscript{35, 37} Subsequently, repulsive forces accumulate as the growing number of positively charged species are pushed closer and closer in space. A fine spray of droplets containing the charged analyte is formed once the repulsive forces become greater than the force of the analyte’s surface tension.\textsuperscript{34, 35, 37} Many theories are proposed for the physical process of droplet formation, including Coulombic fission and solvent evaporation mechanisms.\textsuperscript{37} The theories on droplet formation, along with other details of the ESI process, are extensively reviewed in the literature.\textsuperscript{35, 37}

Certain factors render ESI-MS analysis more challenging, such as the introduction of salts or other impurities into a sample, or when analytes with large differences in ionization energy are present in a mixture.\textsuperscript{21} In these instances, other ionization techniques such as MALDI may be more appropriate to use.\textsuperscript{39}

1.1.3.2 Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) was invented in 1974 by Alan Marshall and co-workers.\textsuperscript{40, 41} FT-ICR MS is one type of mass analyzer that provides highly accurate mass information.\textsuperscript{21, 22, 23, 40, 41, 42} For peptides and their associated PTMs, FT-ICR provides powerful MS analysis capabilities.\textsuperscript{23, 43, 44} The mass error measured between the actual and experimental $m/z$ values for an analyte is typically below 10 parts per million (ppm)
when data is acquired on an FT-ICR MS instrument.\textsuperscript{41} High resolution FT-ICR MS data is achieved by measuring the cyclotron frequency of ions as they pass through a fixed magnetic field,\textsuperscript{40, 41, 42} which is different than the physical contact required for detectors in most other MS instruments.\textsuperscript{45}

Although the MS\textsuperscript{1} data provided by FT-ICR is of high resolution, it is still not sufficient to unequivocally determine a precursor’s composition when two or more possible structures share the same neutral mass.\textsuperscript{27, 28}

1.1.4 Analysis of Protein PTMs by Tandem Mass Spectrometry. The most common method of fragmentation coupled to ESI-MS is collision induced dissociation (CID),\textsuperscript{22} whereas electron transfer dissociation (ETD) is a relatively new method of tandem mass spectrometry used to study peptides.\textsuperscript{46} Both methods result in cleavage along the amino acid backbone of a peptide, though by different mechanisms.\textsuperscript{22} In comparison to CID, which cleaves a peptide between the carbonyl and amine groups of adjacent amino acids,\textsuperscript{22} ETD cleaves non-specifically along the amide bond of amino acids.\textsuperscript{47, 53, 54} ETD allows most labile PTMs to remain intact, while CID produces a signature loss of these modifications.\textsuperscript{47} The two dissociation mechanisms provide complementary results to one another, and each offers distinct advantages in the characterization of PTMs.\textsuperscript{22, 56} That is, the best fragmentation technique is unique to the PTM, as well as the desired information the researcher is trying to obtain.

1.1.4.1 Collision Induced Dissociation Tandem Mass Spectrometry. Collision induced dissociation, also referred to as collisionally activated dissociation (CAD), was first evidenced in the mass spectra of Sir J. J. Thomson.\textsuperscript{57, 58} CID is a type of gas phase ion/neutral pair activation where product ions are generated when a precursor is fragmented indirectly by the transfer of vibrational energy from an inert gas.\textsuperscript{53, 57, 58, 59} The deposit of energy onto an ion
during activation is dependent upon the relative collision energy of the ion/neutral pair that is colliding, which dictates the maximum amount of kinetic energy that is available for transfer of internal energy onto the ion.\textsuperscript{53, 57} One of the useful features of CID is that it is generally universally applicable to analytes; that is, all molecules have a collision cross section.\textsuperscript{57}

The trajectory properties (kinematics) for these ion collisions have been previously described in detail.\textsuperscript{53, 59} Equation 1 shows the available kinetic energy of an ion/neutral pair in which the velocity of the neutral species is broadly considered to be negligible, where $\text{KE}_{\text{COM}}$ indicates the kinetic energy of transfer to the colliding complex, $m_n$ and $m_i$ stand for the mass of the neutral target and precursor ion, respectively, and $\text{KE}$ is the kinetic energy of the ion.\textsuperscript{57} Although the transfer of energy for these reactions is dependent upon a variety of factors, it can generally be classified as high-energy or low-energy CID.\textsuperscript{53, 57}

\textbf{Equation 1.} $\text{KE}_{\text{COM}} = \frac{m_n}{m_n + m_i} \text{ KE}$

The vibrational excitement of analyte ions that occurs in CID results in the dissociation of amide bonds along a peptide’s backbone, and subsequent cleavage between the carbonyl and amine groups of contiguous amino acids.\textsuperscript{22} Product ions that result from this type of cleavage are termed b- and y-type ions.\textsuperscript{22} The product ions detected in CID MS/MS data may then be used to sequence and identify a peptide or protein, and map the location of any chemical or post-translational modification that the protein may possess.\textsuperscript{22, 56}

CID MS/MS is particularly useful in the investigation of protein glycosylation, especially in cases where mass information alone does not support an unambiguous assignment of a glycopeptide’s composition. Specifically, the identity of a glycopeptide bearing multiple
monosaccharide residues can be readily determined from CID experiments using information extracted from the MS/MS data.²⁸

1.1.4.2 Electron Transfer Dissociation Tandem Mass Spectrometry. During ETD MS/MS, fragmentation on positively charged ions is induced by transferring electrons from a radical anion.⁴⁶ ⁵⁰ Unlike the ion/neutral pair activation that occurs in CID, ion/ion pair activation is the basis for ETD fragmentation.²² ETD Reagents such as fluoranthene provide the radical anions necessary for this transfer of electrons to occur.⁴⁶ ⁵⁰

ETD MS/MS experiments generate peptide-containing product ions by cleaving non-specifically along the amide bond of amino acids.²² In this way, ETD is analogous to electron capture dissociation (ECD). These product ions are referred to as c- and z-type ions.²² ⁴⁷ Consequently, ETD allows labile PTMs to remain intact. This is in direct contrast to CID, which produces a signature loss of labile modifications.²² ⁴⁷ To this end, ETD has proven value in the study of proteins with PTMs difficult to characterize using CID.⁴⁶ These include labile modifications of low molecular weight, such as phosphorylation and O-linked GlcNAcylation.⁴⁷

ETD has also shown to be beneficial in the study of proteins modified by disulfide bonds.⁴⁶ For disulfide-bonded peptides, previous work has revealed that ETD preferentially cleaves the disulfide bond between the two joined peptides,⁴⁸ ⁵⁰ ⁶⁰ ⁶¹ leaving a pattern of characteristic product ions that is different in comparison to peptides containing other post-translational modifications.⁶⁰ While CID is useful for obtaining product ions that produce signature losses of labile modifications, such as those resulting from glycosylation, it generally will not cleave a disulfide bond.⁶⁰ Finally, ETD has been shown to impart more extensive peptide sequence coverage.⁴⁷
1.2 PROTEIN GLYCOSYLATION

1.2.1 Overview of Glycosylation. The addition of monosaccharide residues onto a protein or lipid, known as glycosylation, serves an important function in many cellular signaling and communication events, including those involving host-pathogen interactions.\(^6\),\(^6\),\(^6\),\(^6\) It has long been understood that protein-carbohydrate interactions play a participatory role in many processes affecting disease progression.\(^6\),\(^6\),\(^6\),\(^6\),\(^6\) Furthermore, experimental evidence demonstrates that the identity of the attached glycans change during these events.\(^6\),\(^6\),\(^6\),\(^6\) For example, aberrant glycosylation is often present in individuals experiencing cancer, diabetes, and inflammation.\(^6\),\(^6\),\(^6\),\(^6\),\(^6\) Accordingly, accurate characterization of a glycoprotein’s glycan substituents has been shown to be crucial in the development of potential biomarkers, protein-based vaccine candidates, and pharmaceutical treatments.\(^6\),\(^6\),\(^6\),\(^6\)

1.2.2 Glycosylation Heterogeneity. Unlike DNA replication and protein transcription, glycosylation is a “non-template”-driven process,\(^6\),\(^6\) where the sugar residues form a multitude of arrangements.\(^6\),\(^6\) The monosaccharides that comprise the glycan may be long or short, branched or linear, and linked in a variety of ways, creating a large degree of variability.\(^6\),\(^6\),\(^6\) This heterogeneity is described in two ways: Glycan differences at different sites of attachment (macroheterogeneity), or within the same site (microheterogeneity).\(^6\) The large amount of heterogeneity presents a challenging obstacle to researchers attempting to elucidate structural, as well as compositional, information on a protein’s glycan population, especially when samples are mixtures of proteins.

1.2.3 Types of Protein Glycosylation. Over half of all proteins expressed are predicted to be glycosylated.\(^1\) In addition to the established forms of protein glycosylation, including \(N\)-linked, \(O\)-linked, and \(C\)-linked forms,\(^6\),\(^6\),\(^6\),\(^6\) rarer configurations such as \(S\)-linked
glycosylation, are also being discovered. Although a variety of types exist, the two most common types of glycosylation are N-linked and O-linked.

In N-linked glycosylation, the addition of a glycan may occur at the asparagine residue when the consensus sequence Asn-Xaa-Ser/Thr occurs, where Xaa is any amino acid except proline. The inclusion of this pattern is a fundamental requirement for N-linked glycosylation to occur, though it is not a guarantee that a glycosylation site will be occupied. With O-linked glycosylation, the glycan addition may occur at any Ser or Thr residue within the protein sequence, though a very low percentage of these sites are actually occupied. In contrast to N-linked glycans, O-linked glycans have less defined sequence patterns and may consist of several distinct core arrangements. For these reasons, both the prediction and determination of O-linked glycosylation characteristics have advanced slower than N-linked glycosylation analysis.

1.2.4 Characterization of Protein Glycosylation. A variety of biochemical and instrumental techniques may be used to probe a glycoprotein’s features. To obtain glycan structural information, enzymatic sequencing and carbohydrate-binding protein (lectin) arrays are often employed. Likewise, analyses by more utilitarian methods rooted in the separation of carbohydrates, including capillary electrophoresis (CE), high-performance anion-exchange chromatography (HPAEC), and mass spectrometry, are also common. As it stands, mass spectrometry is considered to be the preferred route of analysis for the identification of protein glycoforms, whether attached or released.

1.3 GLYCOSYLATION ANALYSIS BY MASS SPECTROMETRY

In the study of protein glycosylation, mass spectrometry has shown to be a powerful tool, as successful interrogation of glycan composition and structure has largely been achieved
To simplify assignment of mass spectra and other data, each monosaccharide residue of a glycan is represented by a symbol with a unique combination of color and shape. These symbols, and the abbreviation of each associated sugar, are shown in Table 1.

Table 1. Monosaccharide Residue Symbols and Abbreviations.

<table>
<thead>
<tr>
<th>Monosaccharide</th>
<th>Abbreviation</th>
<th>Mass</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fucose</td>
<td>Fuc</td>
<td>146.0579</td>
<td></td>
</tr>
<tr>
<td>Mannose</td>
<td>Hex</td>
<td>162.0528</td>
<td></td>
</tr>
<tr>
<td>Galactose</td>
<td>Hex</td>
<td>162.0528</td>
<td></td>
</tr>
<tr>
<td>N-Acetylglucosamine</td>
<td>HexNAc</td>
<td>203.0794</td>
<td></td>
</tr>
<tr>
<td>Sialic Acid</td>
<td>Neu5Ac</td>
<td>291.0954</td>
<td></td>
</tr>
</tbody>
</table>

1 Mannose and galactose may also be abbreviated as Man and Gal, respectively. However, typical MS data does not distinguish between isomeric structures; therefore, the more general abbreviation of Hex is often used.

There are two main strategies for elucidating protein glycosylation information using MS techniques: 1) Characterization of a protein’s glycans after they are released from glycoproteins and 2) Characterization of glycopeptides after proteolytic digestion of a glycoprotein. The study of released glycans is particularly useful when rapid analysis of glycan composition is desired. Though N- and O-linked glycan populations can be studied independently through the use of different cleavage procedures, no information on where the individual glycans were attached along the protein is obtained when the glycans are cleaved a priori. In order to obtain glycosylation site-specific information for individual glycoforms, the second method, glycopeptide analysis, which requires digestion of the protein using a protease such as trypsin, is necessary. This method is generally advantageous because it provides information about
both glycan composition and the site of the glycan’s attachment.\textsuperscript{83} Despite the associated challenges, techniques that allow for complete profiling of a peptide’s glycan population have advanced greatly in the past decade, especially with respect to site-specific glycopeptide analysis.

To examine protein glycosylation by either of these techniques, a number of resources, including databases providing information on known glycan structures or site of occupancy, as well as collections of experimental data, are currently available.\textsuperscript{86, 87, 88, 89, 90, 91, 92, 93} For instance, researchers needing to identify occupied N-linked glycosylation sites on a specific protein can access UniProtKB,\textsuperscript{86} while those wanting statistics specific to proteins modified by O-GlcNAc could visit dbOGAP.\textsuperscript{93} Although this repertoire of information is greater for proteins modified by N- and O-linked glycan types, databases that contain entries on C-glycosylated proteins, such as dbPTM,\textsuperscript{87} are available as well. A current list and description of these database resources are provided in Table 2.
1.3.1.1 Characterization of High Resolution MS Oligosaccharide Data. Often, the easiest way to identify a protein’s glycan population is by enzymatically cleaving the glycan substituents and analyzing the monosaccharide residues directly.\textsuperscript{85} N- and O-linked glycans from the same protein can be independently characterized in this manner, as in the method described by Goetz \textit{et al.} where β-elimination is used to release O-linked glycans, which are simultaneously permethylated.\textsuperscript{94} Once cleaved, automated analysis tools to assist in the determination of glycan composition from MS data may be used.

One such tool developed to analyze MS data of glycans is Cartoonist, as described by Goldberg \textit{et al.}\textsuperscript{95} This program works to increase the speed of compositional determination in permethylated N-linked glycans from matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) data through identification and annotation of matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) spectra.\textsuperscript{95} The most likely glycan compositions are selected using precursor mass information.\textsuperscript{95} Cartoonist automatically labels MALDI peaks with cartoons of the most probable oligosaccharide structure, as determined by the program’s algorithm, from a library of 300 generated mammalian N-linked glycans.\textsuperscript{95} Recently, Goldberg \textit{et al.} extended this concept by developing an automated tool for the analysis of O-linked glycans from MS/MS data,\textsuperscript{96} as described below in MS/MS approaches for glycan analysis. To date, neither program is publicly available.

Another software solution useful for the identification of glycans from MS data is SysBioWare, described by Vakhrushev \textit{et al.}\textsuperscript{97} SysBioWare takes raw MS\textsuperscript{1} data that a user uploads and performs baseline adjustment and denoising, wavelet analysis, and peak detection before grouping isotopes of detected peaks.\textsuperscript{97} The isotopic grouping is also performed automatically, which enables the program to deduce monoisotopic \textit{m/z} values and precursor
Table 2. Glycosylation Databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>Link to Database</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProtKB</td>
<td><a href="http://www.uniprot.org/">http://www.uniprot.org/</a></td>
<td>N-glycos.</td>
<td>Contains annotation of N-, O-, and C-linked glycosylation, as well as glycation. Both mammalian and non-mammalian entries are provided.</td>
</tr>
<tr>
<td>dbPTM</td>
<td><a href="http://dbptm.mbc.nctu.edu.tw/">http://dbptm.mbc.nctu.edu.tw/</a></td>
<td>N-glycos.</td>
<td>Contains a combinational repertoire of protein PTMs from other databases, including experimentally obtained data on site of modification.</td>
</tr>
<tr>
<td>GlycomeDB</td>
<td><a href="http://www.glycome-db.org/">http://www.glycome-db.org/</a></td>
<td>N-glycos.</td>
<td>Contains over 30,000 carbohydrate structures from all major taxonomies, representing a variety of glycosylation types.</td>
</tr>
<tr>
<td>O-GlycBase</td>
<td><a href="http://www.cbs.dtu.dk/databases/OGLYCBASE/">http://www.cbs.dtu.dk/databases/OGLYCBASE/</a></td>
<td>O-glycos.</td>
<td>Contains over 2000 entries of protein glycosylation sites, the majority of which are O-linked.</td>
</tr>
<tr>
<td>UniPep</td>
<td><a href="http://www.unipep.org/">http://www.unipep.org/</a></td>
<td>N-glycos.</td>
<td>Contains over 1500 entries of N-linked glycosylation sites found in human proteins.</td>
</tr>
<tr>
<td>dbOGAP</td>
<td><a href="http://cbsh.lombardi.georgetown.edu/hulab/OGAP.html">http://cbsh.lombardi.georgetown.edu/hulab/OGAP.html</a></td>
<td>O-glycos.</td>
<td>Contains over 1100 entries on sites modified by O-GlcNAcylation.</td>
</tr>
</tbody>
</table>

1.3.1 Automated Analysis of Released Glycans. In both MS and MS/MS experiments, glycans are frequently investigated independently of the glycoprotein they comprise. To facilitate the profiling of carbohydrates from either data type, a number of automated analysis tools have been described in the literature.
charge states without the need of manual input by the user. Monosaccharide compositions are then determined by the software on the basis of mass. Currently, the SysBioWare program is being updated to include analysis of MS/MS data for glycans as well. SysBioWare is not freely available to the public at this time.

Similar to SysBioWare is GlycoWorkbench. GlycoWorkbench evaluates glycan compositions (which are proposed by the user) by searching the spectral peak list of user-input MS data for matches between calculated theoretical glycan masses and corresponding m/z values. The GlycanBuilder tool, designed to interface with GlycoWorkbench, enables the drawing of glycan structure representations, with all stereochemical information on the monosaccharides depicted as specified by the user. Both analysis tools, GlycanBuilder and GlycoWorkbench, are available online free of charge, as described in Table 3.

GlycoSpectrumScan, another freely available program, was developed by Deshpande et al. and works to identify N-and O-linked glycoforms using MS1 data. This software is capable of analyzing both singly or multiply charged ions directly from raw data, and accepts the input of both ESI and MALDI spectra. GlycoSpectrumScan also determines the relative abundance of N-and O-linked glycoforms that are identified for each glycosylation site. However, the user must enter the N- and/or O-linked glycan compositions potentially present in the sample, as well as the in silico peptide masses of the digested glycoprotein. GlycoSpectrumScan is available online (see Table 3).

1.3.1.2 MS/MS Approaches for Glycan Characterization. Until recently, when automated software tools and scoring algorithms became available, the identification of accurate glycan or glycopeptide assignments from MS/MS data was a key bottle-neck, due to the need for extensive manual data analysis. STAT, designed by Gaucher et al., is one of the first automated
tools for the determination of glycan composition using tandem MS. STAT is designed for glycans of up to ten monosaccharide residues, and has the ability to quickly analyze relevant N-glycan compositions. STAT also lists the most likely structures in order of probability to provide a ranking system when more than one candidate glycan matches the fragmentation profile of the data being analyzed. Unfortunately, this program is no longer publicly accessible.

An early analysis tool capable of evaluating O-linked glycan fragmentation is the OSCAR algorithm. OSCAR, as developed by Ashline et al., is specifically designed for the annotation of permethylated O-linked oligosaccharides from MS^n data. OSCAR is part of a collection of software tools termed Glyspy, which is not currently accessible to the public. Although innovative, the use of OSCAR is limited to direct infusion experiments, as the software does not effectively process data from LC-MS methods.

A program contemporary to OSCAR and also developed to handle glycan MS/MS data is StrOligo. This instrument-specific program was designed by Ethier et al. for the determination of N-linked glycan structures from matrix-assisted laser desorption/ionization tandem mass spectrometry (MALDI MS/MS) data. In published research, StrOligo successfully assigned the correct glycan structure in 24 out of 28 cases. Although the results of these two programs are promising, neither program is freely accessible online.

Several alternative glycan analysis tools are freely available online. One of the earliest of these was reported by Lohmann et al. in 2004. The authors describe the web tools GlycoFragment and GlycoSearchMS, which were developed for glycan structural determination. The theoretical fragmentation patterns of carbohydrate structures are calculated using GlycoFragment, which displays theoretical b- and y-fragments as well as c-, z-,...
a- and x-ions. The GlycoSearchMS works to analyzes experimental glycan data by comparing it against a library of theoretical spectra from N-linked and O-linked glycan fragmentation entries extracted from SweetDB. The GlycoFragment program has been validated on both N-linked and O-linked glycan classes, and, used in conjunction with GlycoSearchMS, enables researchers to determine the most probable glycan composition according to the information from the combined algorithms. Both GlycoFragment and GlycoSearchMS are freely available. See Table 3 for more information.

Another heavily used, free, online tool for glycoform analysis is GlycoWorkbench, which has shown to be a resourceful tool not only for analysis of MS data, as mentioned previously, but in the identification of glycans from MS/MS data as well. To utilize the glycan fragmentation analysis feature, a user must first input/define the possible glycan compositions and spectral peak list. The software then calculates expected glycan fragmentation and relative m/z values, and annotates peaks of the uploaded data with the most probable identity (shown in red to distinguish it), of all compositions tested. As previously stated, GlycoWorkbench is available for free online.

In addition to the freely available tools mentioned above, several other MS/MS analysis tools for glycans are available to researchers, either for purchase or by special request to the tools’ developers. Two of these are GlyCH and Glyquest. GlyCH was developed by Tang et al. to perform automated interpretation of oligosaccharide tandem mass spectra. The algorithm has a scoring function built in to allow researchers to compare compositions when more than one is determined to be possible. The GlyCH algorithm, which has so far been tested on released N-glycans, is also capable of de novo analysis, providing no more than ten monosaccharide residues comprise the glycan chain. Although not freely accessible online;
the program is available upon request from the authors. More recently, Gao et al. developed Glyquest, an automated analysis program that takes a different approach to determine compositions of intact N-linked glycans. This software utilizes a database in conjunction with an integrated search engine to determine the composition of peptide-attached N-glycans from CID MS/MS data. After the program algorithmically identifies the molecular weight of the protonated peptide within a given spectrum; candidate N-glycan compositions are selected and fragmented in silico to generate a theoretical spectrum that is then compared to the experimental spectrum. The glycan compositions with fragmentation profiles that are most similar to the experimental fragmentation are determined to be the most probable candidates. Glyquest is not freely available to the public.

SimGlycan is another program that can be used to increase throughput of glycan analysis. More information is available online (http://www.premierbiosoft.com/). This commercial tool is useful for determining glycan structures from MS/MS data obtained on many different mass spectrometers, once an acquisition file is converted into mzXML format. A user uploads an MS/MS data file, and the software utilizes a built-in database with theoretical fragmentation profiles of nearly 10,000 glycan structures to provide the most likely structural candidates. One unique feature of SimGlycan is that no filtering of biologically relevant structures is provided, which can be advantageous for identifying novel glycan structures, but disadvantageous in that it returns a user many structures which are not pertinent. However, for purchase programs such as SimGlycan are expensive, which potentially limits their use.

A more recent program developed specifically for the compositional interpretation of O-linked glycan fragmentation is CartoonistTwo, as described by Goldberg et al. CartoonistTwo was designed using CID data acquired on an FTICR-MS, and validated using data from a test set.
of 34 spectra acquired from *Xenopus* egg jelly. Unfortunately, the program is not freely accessible to the public.

A summary of those glycan MS and MS/MS data analysis tools that are currently available online is shown below, in Table 3.

**Table 3.** Online Tools to Facilitate Glycan Characterization from MS and MS/MS Data.

<table>
<thead>
<tr>
<th>MS Analysis Tool</th>
<th>Link to Analysis Tool</th>
<th>Concept and Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlycoWorkbench</td>
<td><a href="http://download.glycoworkbench.org/">http://download.glycoworkbench.org/</a></td>
<td>Identifies and annotates MS and MS/MS data with appropriate glycan compositions or fragments.</td>
</tr>
<tr>
<td>GlycoSpectrumScan</td>
<td><a href="http://www.glycospectrumscan.org">http://www.glycospectrumscan.org</a></td>
<td>Quantitatively identifies <em>N</em>- and <em>O</em>-linked glycoforms within a protein using LC-MS data.</td>
</tr>
<tr>
<td>GlycoFragment</td>
<td><a href="http://www.glycosciences.de/tools/GlycoFragments/fragment.php4">http://www.glycosciences.de/tools/GlycoFragments/fragment.php4</a></td>
<td>Identifies and displays the main product ions expected for oligosaccharide MS/MS data.</td>
</tr>
<tr>
<td>GlycoSearchMS</td>
<td><a href="http://www.glycosciences.de/database/start.php?action=form_ms_search">http://www.glycosciences.de/database/start.php?action=form_ms_search</a></td>
<td>Compares experimental MS/MS data to product ions calculated from an extensive library of <em>N</em>- and <em>O</em>-linked glycans.</td>
</tr>
<tr>
<td>SimGlycan</td>
<td><a href="http://www.premierbiosoft.com/glycan/index.html">http://www.premierbiosoft.com/glycan/index.html</a></td>
<td>Predicts the structure of glycans from MS/MS data by matching spectra to a built-in database.</td>
</tr>
</tbody>
</table>

**1.3.2 Automated Analysis of Glycopeptides.** For researchers performing site-specific glycosylation analysis, the initial step toward accomplishing the characterization of attached glycoforms at unique sites within a digested protein is to identify potential glycosylation sites within that protein. The tools to facilitate this step are described in Table 2. In addition to these, programs that utilize algorithms to predict the likelihood of site-occupancy by examination of the amino acid residues surrounding the potential glycosylation site have also been developed.111, 112.
These prediction tools, along with a description and link to each tool, are provided in Table 4.

Table 4. Glycosylation Site Prediction Tools.

<table>
<thead>
<tr>
<th>Database</th>
<th>Link to Database Prediction Tool</th>
<th>Type</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>O-glycos.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C-glycos.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>O-glycos.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>O-glycos.</td>
<td></td>
</tr>
</tbody>
</table>
1.3.2.1 Experimental Data Requirements. After the resultant glycopeptides are obtained from the proteolytic digest, two types of data are generally used to accurately characterize the identity of a glycopeptide. First, high resolution MS data of the glycopeptide is used to infer possible glycopeptide compositions; second, tandem MS data is acquired to distinguish between isomers and isobars. In Figure 2, a schematic of this work-flow is provided.

**Figure 2.** Flow chart outlining the use of MS and MS/MS data for glycopeptide identification.

1.4 AUTOMATED MS and MS" ANALYSIS OF GLYCOPEPTIDES

1.4.1 N-Linked Glycopeptides. Although N-linked glycoforms share a common core structure, the rest of the glycan follows one of three distinct arrangements. Based on the arrangement pattern, N-linked glycans compositions are classified into three main types, those with: 1) High mannose type glycans 2) Complex type glycans and 3) Hybrid type glycans. This information is useful when deciphering glycopeptide compositions from MS experiments, specifically from CID MS/MS data.
1.4.1.1 *N*-Linked Glycopeptide Characterization from MS Data. A variety of automated and semi-automated analysis tools have been created to aid in the interpretation of *N*-linked glycopeptide MS data. The key objective of these tools is to provide glycopeptide compositions that are consistent with the high resolution MS data. Researchers then typically use MS/MS analysis to determine which of the compositions is correct for each given ion. Three of these tools are accessible to the public: GlycoMod (http://web.expasy.org/glycomod/), GlycoPep DB (http://hexose.chem.ku.edu/glycop.htm), and the previously mentioned GlycoSpectrumScan. GlycoMod, the earliest and most heavily used tool, accepts a protein sequence, possible monosaccharide building blocks, and experimental mass data as inputs, and it calculates all possible glycopeptide compositions that fall within the mass tolerance. One restriction in the capacity of GlycoMod to analyze glycopeptide data is the inability to handle multiply charged precursors.

Programs such as GlycoPep DB and GlycoSpectrumScan were designed to overcome some of the limitations in GlycoMod. GlycoPep DB, developed by Go et al. limits its output by restricting the potential glycans in the glycopeptide to a database of biologically relevant glycoforms that have been previously identified in MS data. It also accepts precursor ions in multiple charge states. The disadvantage of using this approach, however, is that if the glycan in the spectrum is not in the GlycoPep DB database, then the software will not be effective at providing the correct assignment for the peak. GlycoSpectrumScan is a more recent program, developed by Deshpande et al., that also interprets MS data on both *N*- and *O*-linked glycopeptides. Like GlycoPep DB, this program has the ability to handle input for both singly and multiply charged data. GlycoSpectrumScan is described in detail below for *O*-linked MS data analysis. Regardless of which tool is used for assigning the high resolution data, these
assignments must be supported by MS/MS data, to provide high confidence assignments.\(^{27}\)

**1.4.1.2 N-Linked Glycopeptide Characterization from MS/MS Data.** Each common N-linked glycan type (complex, hybrid, or high mannose) has a signature fragmentation profile that is present when a glycopeptide is subjected to MS/MS experiments.\(^ {28, 85}\) These characteristic fragmentation profiles are useful for determining the correct identity of an N-linked glycopeptide when isobaric candidate compositions are possible.\(^ {28}\) However, as manual interpretation of these data are challenging, software is required to speed analysis time.

Of the automated tools used to analyze glycopeptides, many are software expansions of programs that were developed previously to analyze released glycans. One disadvantage of expanding glycan analysis tools to glycopeptides is that these tools generally lack capabilities for analyzing and scoring the peptide component of glycopeptides. SimGlycan is one such example. Available for purchase, SimGlycan has been updated to perform fragmentation analysis for glycopeptides, in addition to glycans.\(^ {109, 110}\) As stated previously, SimGlycan uses a database of over 9,000 glycan structures that could be consistent with the MS/MS data to identify the most appropriate composition for the acquired spectrum.\(^ {110}\) SimGlycan may be purchased online (http://www.premierbiosoft.com/).

Many other publicly available tools to elucidate glycosylation profiles of glycopeptides have emerged out of glycan analysis software. GlycoWorkbench and Glyco-Peakfinder both work to annotate glycan fragmentation in glycopeptide data, although the peptide portion of the glycopeptide must be determined by some means other than the use of these tools.\(^ {99, 122}\) On the positive side, Glyco-Peakfinder is useful for *de novo* calculation and annotation of glycan fragment ions within tandem mass spectra.\(^ {122}\) Users may allow constraints on the oligosaccharide such as size and attachment of other substituents (such as acetate, phosphate, and
sulfate), and the program is capable of annotating multiply charged ions (-4 to +4). Additionally, glycan fragmentation is analyzed across multiple charge states, and across multiple charge carriers (cationic carriers), within the same spectrum.

A completely different approach is used in GlycoPep ID. GlycoPep ID is a web-based tool developed by Go et al. to interpret MS/MS data of glycopeptides and to identify the peptide component of glycopeptides through analysis of expected product ions. The URL to access this program is listed in Table 5. Although this program is useful for identification of the peptide portion of the glycopeptide in complex LC-MS samples, it does not contain a scoring algorithm to identify the most probable glycopeptide match.

Software with the ability to score potential compositions is especially useful to researchers. Often, more than one glycan or glycopeptide composition could correspond to a given spectrum within the accepted range of mass tolerance. Therefore, programs that have a scoring function to evaluate each of those possible matches, and return which of them is the most likely structure, greatly improve the efficiency of glycosylation analysis. For tools that lack this feature, a user must spend time manually determining which of the mathematically possible predictions is the best match for the data.

Some alternative, unique strategies have been developed with the goal of scoring MS/MS data against potential glycopeptide compositions, such as those described using Peptoonist, Medicel Integrator, the Branch-and-Bound algorithm, GlycoMaster, Sweet Substitute, and GlyDB. Unfortunately, none of these programs are currently publicly available.

To address the need for publicly accessible tools specifically designed to interpret and score fragmentation of glycopeptides, GlycoMiner was developed by Ozohanics et al. In the
analysis of 3132 spectra, the software was reported to have found 338 that corresponded to MS/MS data of glycopeptides (versus peptides). Designed using quadrupole time-of-flight (Q-TOF) data, this program is capable of assigning glycopeptide compositions when both the peptide and glycan components portions are unknown. However, GlycoMiner is only capable of performing compositional analysis when the spectra are of good quality. The program fails when spectral quality is low, as evidenced by the software’s identification of glycan composition in only 196/338 glycopeptide spectra. Although this tool is a great advancement towards automated interpretation of glycopeptide MS/MS data, GlycoMiner often generates multiple plausible compositions and fails to rank the correct glycopeptide as the top candidate. In addition, the program requires spectra containing a low S/N, as well as the presence of glycopeptide oxonium marker ions, which are not typically present in data collected on ion trap instruments. Available online, GlycoMiner is free to download and use; see Table 5.

Similar to GlycoMiner, GlycoPeptide Search (GPS) is a recently developed program by Chandler et al. for the determination of glycopeptide composition from CID data. Designed for purified glycoprotein samples analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS), GPS utilizes GlycomeDB, a glycan database in conjunction with the peptide file, which is supplied by the user, to produce an Excel file of glycopeptide matches based on fragmentation evidence. To generate the peptide-glycan pairs, GPS must find both low mass oxonium, and N-glycan core-containing, product ions. GPS is freely available online, as well. For further information, see Table 5.

The targeted MS/MS approach utilizing the computational tool GlypID recently described by Wu et al. aims to characterize N-linked glycopeptides through the combined use of MS\(^1\) and MS\(^2\) information extracted from LC-MS/MS experiments. One of the benefits to the
method is that no prior knowledge of the potential glycosylation or identity of the glycopeptide is necessary. Instead, GlypID assigns a cluster of glycopeptides in the “same family” (microheterogeneities) based on observed mass. In addition, the approach utilizes an isotope deconvolution algorithm to assign ion charges along with monoisotopic ions. This information is then added to the inclusion list of “prioritized precursor ions” for the MS/MS analysis that follows. Next, the resultant CID data is searched for the longest series of glycosidic bond cleavage series. These product ions are used to determine the oligosaccharide sequence tag, which is used to verify whether or not the spectrum is from a glycopeptide. A score is assigned to the CID spectrum based on this sequence tag. MS data is used to evaluate and score the relative probability of a glycopeptide by examining the clusters of peptide glycoforms, or those glycopeptides with the same peptide backbone that co-elute within a specific time range. The glycoform is then identified using the mass of the attached N-linked glycan, though the most current version of GlypID allows the entry of user-defined glycan compositions as well. A limitation to the program is that when low resolution data is used, there is a significant increase in the number of false-positive identifications of glycopeptide microheterogeneities within a cluster. Although the new targeted MS/MS approach has been optimized for FT MS instrumentation and data, the original GlypID algorithm was designed using LC-MS ion trap data. A publicly accessible version of the computational tool is currently available online, free of charge to users (see Table 5).

Mayampurath et al. recently modified the GlypID algorithm with a scoring function that works to determine glycopeptide composition from high-energy C-trap dissociation (HCD) MS/MS data. The new software tool, GlypID 2.0, uses high resolution MS data along with CID and HCD scan information to improve the accuracy of N-linked glycopeptide
identification. Like the original GlypID, GlypID 2.0 can also score CID spectra independently on MS systems that do not contain the HCD instrument option. GlypID 2.0 is freely available to download, as listed in Table 5.

Woodin et al. have also developed a freely accessible web-based tool, GlycoPep Grader (GPG), to assign glycopeptide composition from MS/MS data in an automated fashion. This tool is specifically designed for data collected in an ion trap mass spectrometer, and it features a novel algorithm that enables users to identify the correct glycopeptide composition from a pool of candidate compositions of the same nominal mass.

GPG utilizes the MS/MS data by calculating, scoring, and searching for the expected product ions of potential glycopeptide candidate compositions. The algorithm scores the glycopeptide candidate composition through detection of two types of product ions: 1) Ions that contain the peptide portion and some portion of the pentasaccharide core, or [peptide + core component] ions, and 2) Ions formed via neutral loss of monosaccharide residues from the precursor ion, or [precursor – monosaccharide] ions. The algorithm that powers GPG has been shown to assign the correct glycopeptide candidate after performing the MS/MS peak list search with a very high degree of accuracy.

One advantage to the algorithm behind GPG is that the precursor ion’s charge state is included in the input data, so all product ions can be searched for within their appropriate charge state. Secondly, no spectral transformation (to singly charged ions) needs to be performed prior to using the program, as GPG automatically searches for product ions in a charge-specific fashion, bypassing the need for additional processing software. A disadvantage of the program is that the user must utilize a separate program, such as GlycoMod, to obtain potential matches for the high resolution MS data, prior to assigning the MS² data with GPG. GPG can be found
online, and is free to use.

A summary of programs that assist in N-linked glycopeptide characterization from MS/MS data is listed in Table 5.
Table 5. Freely Available N-linked Glycopeptide Analysis Tools.

<table>
<thead>
<tr>
<th>Database</th>
<th>Link to Database</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlycoMod</td>
<td><a href="http://web.expasy.org/glycomod/">http://web.expasy.org/glycomod/</a></td>
<td>GlycoMod determines potential glycopeptide compositions, on the basis of mass information, from MS data.</td>
</tr>
<tr>
<td>GlycoPep DB</td>
<td><a href="http://hexose.chem.ku.edu/glycop.htm">http://hexose.chem.ku.edu/glycop.htm</a></td>
<td>GlycoPep DB deduces possible biologically relevant glycan compositions from MS data of glycopeptides with a “smart search”</td>
</tr>
<tr>
<td>GlycoSpectrumScan</td>
<td><a href="http://www.glycospectrumscan.org">http://www.glycospectrumscan.org</a>.</td>
<td>GlycoSpectrumScan searches LC-MS data to identify glycopeptides and determine glycoform location.</td>
</tr>
<tr>
<td>GlycoWorkbench</td>
<td><a href="http://download.glycoworkbench.org/">http://download.glycoworkbench.org/</a></td>
<td>GlycoWorkbench annotates glycopeptide MS/MS data through fragmentation analysis and scoring of only the glycan portion.</td>
</tr>
<tr>
<td>Glyco-Peakfinder</td>
<td><a href="http://glyco-peakfinder.org/">http://glyco-peakfinder.org/</a></td>
<td>Glyco-Peakfinder performs de novo analysis of glycopeptides, after a peptide sequence is input by a user, using glycan fragmentation profiling.</td>
</tr>
<tr>
<td>GlycoPep ID</td>
<td><a href="http://hexose.chem.ku.edu/predictiontable2.php">http://hexose.chem.ku.edu/predictiontable2.php</a></td>
<td>GlycoPep ID analyzes MS/MS glycopeptide data from complex mixtures by identifying the peptide portion based on expected product ions.</td>
</tr>
<tr>
<td>GlycoMiner</td>
<td><a href="http://www.chemres.hu/ms/glycominer/tutorial.html">http://www.chemres.hu/ms/glycominer/tutorial.html</a></td>
<td>GlycoMiner identifies glycopeptides in qTOF MS/MS data, and assigns composition for quality spectra containing specific marker ions.</td>
</tr>
<tr>
<td>GPS</td>
<td><a href="http://edwardslab.bmcb.georgetown.edu/software/GlycoPeptideSearch.html">http://edwardslab.bmcb.georgetown.edu/software/GlycoPeptideSearch.html</a></td>
<td>GPS generates glycopeptide compositions, utilizing a glycan database, after searching and matching LC-MS/MS data of purified proteins.</td>
</tr>
<tr>
<td>GlypID</td>
<td><a href="http://www.cbs.dtu.dk/services/DictyOGlyc/">http://www.cbs.dtu.dk/services/DictyOGlyc/</a></td>
<td>GlypID identifies glycopeptides from LC-MS/MS experiments using a combination of MS\textsuperscript{1} and MS\textsuperscript{2} data.</td>
</tr>
<tr>
<td>GlypID 2.0</td>
<td><a href="http://mendel.informatics.indiana.edu/~chuyu/glypID/software.html">http://mendel.informatics.indiana.edu/~chuyu/glypID/software.html</a></td>
<td>GlypID 2.0 uses CID and HCD MS/MS data to deduce monosaccharide composition, as well as glycan type and location, for N-glycopeptides.</td>
</tr>
<tr>
<td>GPG</td>
<td><a href="http://glycopro.chem.ku.edu/GPGHome.php">http://glycopro.chem.ku.edu/GPGHome.php</a></td>
<td>GPG scores glycopeptide candidates after searching MS/MS data for each candidate’s predicted product ions.</td>
</tr>
</tbody>
</table>

1.4.2 O-Linked Glycopeptides. The analysis of O-linked glycoforms is particularly challenging, as no single consensus sequence exists to predict the site of glycan attachment.\textsuperscript{68, 80}
Further adding to the difficulty of analysis, factors that affect the efficiency of glycosylation at N-linked sites are different than those affecting O-glycosylation efficiency. For example, the presence of aromatic residues near an O-linked site inhibits glycosylation; whereas the presence of an aromatic residue near an N-linked site increases the likelihood of site-occupancy.\textsuperscript{82}

1.4.2.1 Mucin-Type O-Linked Glycosylation. The most prevalent form of O-linked glycosylation to occur in eukaryotic organisms is mucin-type O-glycosylation, which occurs where glycans are attached to a protein by the addition of α-N-acetylgalactosamine (GalNAc) residues to the hydroxyl group of Ser/Thr side chains (commonly referred to as the Tn antigen).\textsuperscript{63, 68} Though still in the infancy stage, analysis tools have recently been created to assist researchers in the determination of O-linked glycoforms, many of which are mucin in type, from MS data.

1.4.2.2 O-Linked Glycopeptide Characterization from MS Data. Recently, Deshpande et al. advanced the MS data analysis of N- and O-linked glycopeptides with the advent of the GlycoSpectrumScan program.\textsuperscript{101} GlycoSpectrumScan is designed to analyze LC-MS data of intact glycopeptides from proteolytic digests.\textsuperscript{101} The program utilizes MS\textsuperscript{1} data to determine glycopeptide composition, along with the relative distribution of glycoforms at each of the sites.\textsuperscript{101} In addition, the algorithm behind the program offers a few distinct advantages in that it handles multiply charged ions, making it amenable to both MALDI and ESI data, and is currently freely available online (www.glycospectrumscan.org.).\textsuperscript{101}

GlycoX and GlycoMod, described earlier in the analysis of N-linked glycopeptides, are capable of O-linked glycopeptide data interpretation as well.\textsuperscript{120, 136} Unlike GlycoMod (http://web.expasy.org/glycomod/), GlycoX is not publicly available, though it is available upon request from the authors.\textsuperscript{136} GlycoWorkbench, also described previously, performs automation
of O-linked glycopeptide MS\(^1\) data to elucidate the most likely composition from an experimental peak list in the same manner as for N-linked glycopeptide spectra.\(^99\) GlycoWorkbench is freely available online (http://download.glycoworkbench.org/).

### 1.4.2.3 O-Linked Glycopeptide Characterization from MS/MS Data.

Currently, there is no freely available stand-alone program designed to automate the analysis of O-linked glycopeptide CID MS/MS data through evaluation of both unknown portions of a glycopeptide, the peptide and glycan. The GlycoWorkbench program is capable of annotating glycans in CID fragmentation data of glycopeptides.\(^99\) However, as described for the MS/MS characterization of N-linked glycopeptides, the identity of the peptide portion must already be known, as GlycoWorkbench solely evaluates the fragmentation of the glycan-containing portion of a glycopeptide.\(^99\)

There are promising advances being made in the compositional determination of glycopeptides using ETD fragmentation techniques,\(^72,84\) or a combination of CID and ETD, particularly in the study of O-linked species.\(^137\) A recent method described by Darula et al. in which MS\(^1\), CID, and ETD data are used in conjunction with Protein Prospector v5.3 for the identification of SA\(_{1,10}\)GalGalNAc-containing O-linked glycopeptides enriched from bovine serum, demonstrates the potential for automated analysis through a combination of these techniques and database searches.\(^137\) However, this process is only semi-automated, and restricted to samples containing simple carbohydrate structures.\(^137\) Hopefully, the compositional information gained between the two complementary fragmentation methods of CID and ETD will enable researchers to gain insight into creating automated programs to speed the analysis of O-linked MS/MS glycopeptide data as well.
1.5 PROTEIN DISULFIDE BOND FORMATION

1.5.1 Overview of Disulfide Bonds. Protein disulfide bonds result when two cysteine residues are covalently joined through oxidation.\textsuperscript{1, 4, 7} The formation of native disulfide bonds in a protein is necessary to achieve proper folding, stability and conformation.\textsuperscript{1, 8, 9, 14, 15} In addition, disulfide bonds directly contribute to many biological functions by participating in cellular regulation or catalysis events.\textsuperscript{1, 9, 15} Furthermore, dysregulation of enzymes that are involved in the formation of disulfide bonds have been reported in some diseases.\textsuperscript{15} Aberrations of disulfide bond structure for proteins showing altered expression or activity have also been reported in damage caused by oxidative stress events.\textsuperscript{5, 15}

The formation of disulfide bonds within a protein is paramount to the production of functional biopharmaceuticals. Often, treatments are designed using proteins containing modifications, and utilize recombinant protein technology.\textsuperscript{6, 13, 15, 138, 139} For many of these, \textit{E. coli} has been used to express the associated recombinant proteins. This is problematic, as disulfide bond scrambling and misfolding are often reported for proteins produced by \textit{E. coli}.\textsuperscript{139}

The folding of the final protein product is critical to its biological function. Consequently, there is a need to obtain timely, accurate and concise data so the quality of protein therapeutics can be assessed. To this end, thorough scrutiny into the pattern and integrity of each associated protein’s disulfide bond arrangements become necessary. According to Trivedi \textit{et al.}, the demand for this characterization is rapidly increasing,\textsuperscript{14} due to a rising trend in the use of protein-based drugs.

1.5.2 Types of Disulfide Bonding. There are two main types of protein disulfide arrangements that are commonly found in proteins and peptides: Interchain and intrachain type bonds.\textsuperscript{18} Interchain type bonds occur when the disulfide bond is comprised of cysteine residues
from more than one protein or protein subunit, and intrachain type bonds occur when the cysteine residues are within the same protein or protein subunit. Figure 3 provides a visualization of interchain disulfide bonding. In Fig. 3, two discrete subunits of a single protein are shown with their participating thiols orientated in close proximity to form an interchain disulfide bond upon oxidation of each cysteine’s sulfhydryl group.

Figure 3. Interchain disulfide bonding between two subunits from the same protein. The disulfide bond linking the oxidized cysteine residues is shown in red.

For peptides, interchain type bonding refers to disulfides that form between two or more unique peptides, and intrachain type bonding indicates that the cysteine residues forming the disulfide reside along the amino acid sequence of a single peptide.\textsuperscript{18, 50, 140}

1.5.3 Characterization of Protein Disulfide Bonds. To investigate the formation and integrity of protein disulfide bonds, a variety of spectroscopic and biochemical techniques may be exploited. Before the advent of appropriate MS instrumentation, NMR and crystallography were commonly used to characterize disulfide connectivity.\textsuperscript{141, 142} Unfortunately, both NMR and crystallography experiments require large amounts of high purity samples, which are often difficult to obtain. Edman degradation is another traditional method used to investigate protein disulfide bond patterns; however, this method requires that samples be of ultra-high purity.\textsuperscript{18}
1.5.4 Disulfide-Bonded Peptide Analysis by Mass Spectrometry. In the characterization of peptides containing intact disulfide bonds, a variety of MS methods have shown to be effective, including MALDI-TOF and ESI-FT-ICR. However, as ETD is a relatively new fragmentation technique, the availability of mass analyzers that can be coupled to ETD is not near the number that can be coupled to CID. To date, limited instruments are equipped with the ability to perform ETD MS/MS analysis. Of these MS systems, not all provide high resolution MS scans. With low resolution mass spectrometers, it is often necessary to assign charge state independent of the MS data. In these instances, the charge state of disulfide-bonded precursors needs to be determined using a different approach.

1.5.5 Disulfide-Bonded Peptide Analysis by Tandem Mass Spectrometry. MS analysis of peptides containing intact disulfide bonds has been advanced by the advent of recently developed MS/MS fragmentation techniques, such as infrared multiphoton dissociation (IRMPD), ECD, and ETD. Although all three techniques have proven to be powerful tools for profiling species that are difficult to analyze by CID, IRMPD and ECD are much more costly than ETD. As mentioned previously, disulfide bonds are readily cleaved in ETD MS/MS. This is a particularly informative characteristic of disulfide-bonded peptide ETD data, as detection of the product ions arising from the individual peptide chains have proven useful for identifying the composition of the intact disulfide-bonded precursor.

1.5.5.1 Automated MS/MS Data Analysis of Disulfide-Bonded Peptides. ETD fragmentation was first described in 2004, and computational tools to assist in the analysis of resultant MS/MS data have not yet advanced to the level seen for CID MS/MS data. In particular, analysis tools that work to determine the charge state of precursor ions in low resolution ETD fragmentation data, where charge state assignments are not apparent from
isotopic distribution, are needed. Although a few automated peptide ETD MS/MS analysis programs have been created, these software were not developed or tested using disulfide-bonded peptides, and are either not freely available or difficult for persons not trained in the use of complex software to use.\textsuperscript{149, 150, 151}

1.6 CONCLUDING REMARKS

Mass spectrometry is often the method of choice for elucidating protein post-translational modifications. The generation of automated MS and MS/MS analysis tools to assist in the characterization of PTMs is emerging as an effort to facilitate more rapid analysis of data collected on proteins and peptides that contain them. Specifically, analytical approaches and automated tool development for the investigation of protein glycosylation and disulfide bond formation, two of the most common PTMs, are discussed herein.

For the study of protein glycosylation, there are two main approaches used by researchers: Glycan analysis and glycopeptide analysis. The least challenging mode of analysis is to release the glycans from a glycoprotein and analyze them independently. However, the most informative approach is to utilize a protease and cleave the glycoprotein into glycopeptides, thereby retaining information on where each glycan is attached within the protein sequence.

Similar to glycopeptide analysis, MS/MS experiments can be used to identify and map the location of disulfide bonds on a protein after it has been enzymatically cleaved into peptides. A key difference for proteins containing disulfide bonds is that the cysteine residues are not reduced prior to proteolytic digestion, in order to retain the disulfide linkages.

Current research shows that although progress has been made in the development of software for peptides containing glycans or disulfide bonds, there are still crucial deficiencies that must be overcome before MS analysis of these and other PTMs is fully automated.
1.7 ACKNOWLEDGEMENTS

The author acknowledges financial support from the National Institutes of Health (RO1RR026061) and an NSF Career award (0645120) to H.D., an NSF Fellowship (DGE-0742523) and Pfizer Award to C.W., and a Seo Scholarship to M.M. This includes support to publish the manuscript incorporated into Chapter 1 of this Dissertation.

The author also recognizes the contribution of co-authors in making the publication of the manuscript mentioned herein possible: Morgan Maxon for her time spent gathering information on a number of glycosylation databases and analysis tools, and Heather Desaire for her intellectual input and advice.
1.8 SUMMARY OF SUBSEQUENT CHAPTERS

Chapter 2 describes a set of fragmentation rules that predict product ion formation in the tandem mass spectra of peptides post-translationally modified by glycosylation. These rules were developed after extensive analysis of the dissociation patterns detected in experimental N-CID MS/MS data collected on these complex species. Prior to MS analysis, model glycoproteins were digested by trypsin to yield glycopeptides comprised of various N-linked glycan arrangements. The fragmentation rules developed from these studies are applicable to all N-linked glycopeptides, regardless of the type of monosaccharide residues that comprise the glycans. Finally, these rules were used to devise an algorithm that would be the basis for MS/MS data analysis software.

Chapter 3 encompasses the development and testing of glycopeptide software, GlycoPep Grader (GPG). The GPG software incorporates the original algorithm that was created from the CID studies on N-linked glycopeptides. Specifically, the analysis tool evaluates MS/MS data for the presence or absence of predicted products to elucidate N-linked glycopeptide composition. GPG was first tested on the collection of CID spectra from the fragmentation studies (training data set) before it was applied to a protein that was not part of the original algorithm design (validation data set). For both data sets, GPG detected the correct composition from a pool of candidate glycopeptides of nearly identical mass (actual and decoy) in every test performed.

Chapter 4 describes an MS/MS analysis tool for the determination of precursor charge state from peptides containing another common PTM, the inclusion of disulfide bonds. Proteins with a variety of disulfide bond patterns were digested using a protease, in the absence of a reducing agent, in order to yield disulfide-bonded peptides. These peptides were then analyzed by ETD MS/MS to develop a method for the determination of precursor charge state directly.
from the fragmentation data. Finally, the devised computational approach was automated with 2 straight-forward and easy-to-use computational tools that a user may easily reproduce in Excel.

Chapter 5 outlines future updates to the GPG software with the goal of improving the score separation between the correct, or actual glycopeptide composition, and the decoy candidates. After scoring hundreds of CID spectra, alternative fragmentation patterns were noted for complex/hybrid type glycopeptides bearing labile terminal residues. In addition, other proposed updates to GPG scoring for all complex/hybrid type glycopeptides are also discussed.
1.9 REFERENCES


(57) Wells, J. M.; McLuckey, S. A. Collision-induced dissociation (CID) of peptides and


(71) Lefebvre, T.; Dehennault, V.; Guinez, C.; Olivier, S.; Drougart, L.; Mir, A. M.; Mortuaire, M.; Vercoutter-Edouart, A. S.; Michalski, J. C. Dysregulation of the nutrient/stress sensor O-


(85) Nwosu, C. C.; Seipert, R. R.; Strum, J. S.; Hua, S. S.; An, H. J.; Zivkovic, A. M.; German,


(98) Vakhrushev, S. Y.; Dadimov, D.; Peter-Katalinić, J. SysBioWare: Structure assignment tool


(137) Darula, Z.; Chalkley, R. J.; Baker, P.; Burlingame, A. L.; Medzihradszky, K. F. Mass spectrometric analysis, automated identification and complete annotation of O-linked


(150) Sadygov, R.; Hao, Z.; Huhmer, A. F. R. Charger: Combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer

CHAPTER 2

COLLISION INDUCED DISSOCIATION BEHAVIOR OF N-LINKED GLYCOPEPTIDES

The work described in Chapter 2 encompasses an original (first author) publication:


ABSTRACT

In order to accurately determine glycopeptide composition using mass spectrometry (MS), fragmentation information is necessary. For N-linked glycopeptide precursor ions fragmented by collision induced dissociation tandem mass spectrometry (CID MS/MS), the dissociation profiles obtained are uniquely correlated to a glycopeptide’s glycan substituent. This information, along with precursor m/z, allows composition to be deduced with high accuracy. However, manual interpretation of these spectra is both challenging and laborious, and limited programs exist to assist in the characterization of glycopeptides from CID data. Developing a set of fragmentation rules is paramount toward designing the necessary algorithms to successfully automate this MS/MS analysis.

In this work, experimental MS studies on N-linked glycopeptides were performed in order to create a set of fragmentation rules to act as the basis of a novel glycopeptide MS/MS scoring algorithm. Liquid-chromatography tandem mass spectrometry (LC-MS/MS) was done on glycopeptides generated from tryptic digestion of RNase B, asialofetuin, and transferrin to determine common product ions for the different types of N-linked glycopeptides that exist. Resultant CID spectra, along with the large body of literature on MS/MS data of glycopeptides, were then used to define a set of fragmentation rules applicable to all N-linked glycopeptides,
regardless of type. These rules incorporate differences in fragmentation that were found to be present for different glycopeptides, and dependent on the monosaccharide arrangement of the glycan substituent. Next, the set of fragmentation rules were incorporated into a novel scoring algorithm that deciphers glycopeptide composition from MS/MS data. Specifically, the algorithm searches a CID spectrum for characteristic product ions predicted to be present for specified N-linked glycopeptide candidates and identifies the most likely composition of both the peptide and the attached glycan.
2.1 INTRODUCTION

In the human body alone, over half of all proteins expressed are predicted to be glycosylated.\textsuperscript{1} Cellular communication events such as signaling, targeting and transport are also known to be proudly impacted, even dependent, upon the types of glycosylation present for a given protein.\textsuperscript{2, 3, 4, 5, 6, 7, 8, 9} It should come as no surprise then that a variety of adverse physiological conditions, such as inflammation and diabetes, and numerous disease states, including cancer, typically present alongside an aberration of glycans on those proteins affected.\textsuperscript{10, 7, 8, 9, 11, 12} In order to develop effective pharmaceuticals for the treatment of those afflicted with such disorders, accurate glycan profiling of those involved glycoproteins is necessary. One of the most common ways to accomplish this is through the use of mass spectrometry (MS) experiments.\textsuperscript{13, 14, 15, 16, 17, 18}

The use of MS for the interrogation of protein glycosylation is accomplished by two main routes: 1) Analysis of released glycans, and 2) Analysis of glycopeptides.\textsuperscript{13, 16} Although insight on glycan composition and relative abundance is achieved using either approach, glycopeptide analysis is most advantageous in that it provides location evidence for each individual glycan residing on a peptide.\textsuperscript{16, 17} In the study of glycopeptides, the use of tandem mass spectrometry (MS/MS) is especially important. Information from MS/MS experiments allow a glycopeptide’s composition to be determined in cases where MS\textsuperscript{1} data alone is not sufficient to do so: When the experimental mass of the precursor ion, within a specified error range, correlates to more than one possible structure.\textsuperscript{19}

Several distinctive features render MS/MS by collision induced dissociation (CID) amenable to glycopeptide analysis. One is that CID MS/MS permits glycopeptide data to be readily distinguished within a protease digest, even though glycopeptides are generally present in
low concentration as compared to peptides.\textsuperscript{20} This is due to the characteristic marker ions that are detected in, and diagnostic of, spectra pertaining to them. These low mass oxonium ions ($m/z$ 204, $m/z$ 163, $m/z$ 292, $m/z$ 366 and $m/z$ 657) allow for the identification of glycopeptide spectra even when complex samples are considered, and also serve as an indicator for which monosaccharide residues comprise the attached glycan.\textsuperscript{20, 21, 22, 23, 24} The existence of these marker ions is indicative of which terminal monosaccharide residues comprise the carbohydrate portion, but not adequate to decipher overall glycopeptide composition. To this end, tedious evaluation of the entire CID spectrum for product ions encompassing multiple aspects of a glycopeptide precursor’s fragmentation is necessary.

Glycopeptides have been shown to dissociate during CID MS/MS on the basis of their attached glycan arrangement.\textsuperscript{20, 24, 25} It is through the study of these distinct fragmentation profiles that allow a glycopeptide to be accurately correlated to their tandem mass spectra. However, no set of comprehensive fragmentation rules have been reported for glycopeptide data thus far. As a result, researchers must rely on careful manual analysis in order to assign glycopeptide composition to a given CID spectrum. Although this analysis is now routinely performed, it remains a complex and difficult task, as two unknowns that must be identified are present: The peptide portion, and the glycan portion.\textsuperscript{16}

Due to these challenges, analysis programs to aid in the interpretation of glycopeptide MS/MS data are limited. Current research efforts toward improved automation are discussed in Chapter 3 (Introduction) of this dissertation. In order to develop effective algorithms to power these automated glycopeptide tools, a set of rules that accurately describes their fragmentation profiles must be developed. These rules must be applicable to all $N$-linked glycopeptides, and therefore incorporate the unique properties for each of the potential carbohydrate substituents.
that comprise them.

Herein, we describe CID MS/MS experiments performed on glycopeptides of various glycan types. After a large collection of spectra was obtained, they were extensively analyzed in order to develop a set of fragmentation rules to develop an initial algorithm to expedite the analysis of CID data collected for any N-linked glycopeptide. Finally, these fragmentation rules were utilized to construct an initial algorithm to serve as the basis for the automated computer analysis tool described in Chapter 3 of this dissertation.

2.2 EXPERIMENTAL

2.2.1 Materials and Reagents. Bovine asialofetuin, bovine ribonuclease B (RNase B), human apo-transferrin (transferrin), urea, dithiothreitol (DTT), iodoacetamide (IAM), formic acid, acetic acid, Sepharose® CL-4B, HPLC grade ethanol, and HPLC grade 1-butanol were purchased from Sigma Aldrich (St. Louis, MO). HPLC grade methanol (CH₃ OH) and HPLC grade acetonitrile (CH₃ CN) were purchased from Fisher Scientific (Fairlawn, NJ). Ammonium bicarbonate (NH₄HCO₃) was purchased from Fluka (Milwaukee, WI) and sequencing grade modified trypsin was from purchased Promega (Madison, WI). Ultrapure water was obtained from a Millipore Direct-Q® UV 3 system (Billerica, MA) with a resistance greater than 18 MΩ.

2.2.2 Preparation of RNase B, Asialofetuin, and Transferrin Glycopeptides. To obtain glycopeptide samples, approximately 300 µg of each protein was dissolved in 50 mM NH₄HCO₃ (pH 8.0) containing 4-6 M urea for denaturation. Disulfide bonds were reduced by the addition of 15 mM DTT and incubation at room temperature for 1 hr. Samples were then alkylated by allowing 25 mM IAM to react with the reduced glycoproteins at room temperature in the dark, for an additional period of 1 hr. The alkylation reaction was quenched through the addition of 40 mM DTT. Next, trypsin was added in a 1:30 (w/w) protease to protein ratio and
incubated at 37 °C for 18 hr. The protease digestion was stopped by the addition of 1 µL concentrated acetic acid for every 100 µL solution. After digestion, RNase B and asialofetuin samples were subjected to glycopeptide enrichment by an in-solution extraction method as described by Rebecchi et al. These samples were then analyzed by direct infusion, as described below. The transferrin sample was not enriched, as it was analyzed by LC-MS, also described below.

2.2.3 Direct Injection Mass Spectrometry. RNase B and asialofetuin samples were reconstituted after glycopeptide enrichment using solvent consisting of 1:1 (v/v) ultrapure water/methanol in 0.5 % acetic acid, to a final concentration of 10 µM immediately prior to direct injection of the glycopeptide samples onto an ESI-LIT-FTICR mass spectrometer (ThermoScientific, San Jose, CA) containing a 7 Tesla actively shielded magnet. Samples were injected at a flow rate of 1 µL/min, and data was collected in positive ion mode. Optimization of the spray voltage was performed to achieve maximum signal. The carrier gas, N₂, was set to 10 psi and the capillary temperature was set to 200 °C. A 2 Da isolation window was used to select precursor ions for MS/MS experiments. Activation time was set to 30 ms, activation qₑ was set to 0.250, and activation energy was set to 30 %, as defined by the instrument software. Thirty scans, each with 10 microscans, were averaged during the collection of MS/MS data.

2.2.4 Liquid Chromatography and Mass Spectrometry. Transferrin glycopeptides were subjected to LC-MS experiments. High resolution MS and MS/MS data were acquired on an ESI-LIT-FTICR-MS (ThermoScientific, San Jose, CA) containing a 7 Tesla actively shielded magnet. The mass spectrometer was directly coupled to a Dionex UltiMate capillary LC system (Sunnyvale, CA) equipped with a FAMOS well plate autosampler. The mobile phase for solvent A was comprised of 99.9 % H₂O + 0.1 % formic acid and the mobile phase for solvent B
consisted of 99.9 % CH\textsubscript{3}CN + 0.1 % formic acid. Transferrin glycopeptides (5 µL at 3 µg/µL) were injected onto a C18 column (300 µm i.d. x 5 cm, 3 µm particle size, CVC MicroTech, Fontana, CA). The flow rate was set to 5 µL/min. Solvent conditions were as follows: 5 min at 5 % B, a 50 min. linear increase to 40 % B, a 10 min linear increase to 90 % B, 10 min. at 90 % B, and re-equilibration of the column. To prevent sample carryover, a 30 min wash cycle followed by a blank run was performed between each sample. The capillary offset voltage was 47 V, capillary temperature was 200 °C, and the spray voltage on the ESI source was set to 2.8 kV. Mass spectrometry data were collected in a data dependent manner. The five most intense ions were selected for collision induced dissociation (CID) in the linear ion trap using 30 % collision energy, and a dynamic exclusion window of 3 min was included.

2.2.5 Manual Data Analysis. To identify the glycopeptides from these samples in the MS data, a prediction table of theoretical m/z values corresponding to glycopeptide compositions for each of the three proteins was prepared. The amino acid sequences from RNase B, asialofetuin, and transferrin were obtained from Uniprot (www.uniprot.org) and their sequences were imported into Protein Prospector (http://prospector.ucsf.edu/prospector/mshome.htm) where tryptic peptides containing Cys residues were modified with carbamidomethylation, and a theoretical tryptic digest was performed to consider up to two tryptic miscleavages. The masses of the peptides that contained potential N-linked glycosylation sites were added to the masses of the known glycan compositions for each glycosylation site, in order to obtain glycopeptide masses. These masses were converted into m/z values corresponding to the glycopeptides in multiple charge states. The MS/MS data for RNase B, asialofetuin, and transferrin were then searched to identify spectra that corresponded to the correct m/z value for a given glycopeptide composition. The MS/MS data were carefully (manually) evaluated, to verify the glycopeptide
assignment.

2.3 RESULTS AND DISCUSSION

*N*-linked glycopeptides of different types have been shown to generate unique dissociation profiles during MS/MS experiments.\(^{25}\) In order to develop a set of rules that can be used to predict the expected product ions applicable these various compositions, three model glycoproteins with differing *N*-glycan moieties were utilized during these experimental studies. These include the well-characterized RNase B, asialofetuin, and transferrin glycoproteins, the properties of which are shown in Table 1.

Table 1. Glycopeptides Analyzed by CID MS/MS.

<table>
<thead>
<tr>
<th>Glycoprotein</th>
<th>Mass (Da)</th>
<th>Length (AA)</th>
<th># <em>N</em>-glycan Sites&lt;sup&gt;†&lt;/sup&gt;</th>
<th><em>N</em>-Glycan Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNase B</td>
<td>16,461</td>
<td>150</td>
<td>1</td>
<td>High Mannose</td>
</tr>
<tr>
<td>Asialofetuin</td>
<td>38,419</td>
<td>359</td>
<td>3</td>
<td>Complex</td>
</tr>
<tr>
<td>Transferrin</td>
<td>77,064</td>
<td>698</td>
<td>2</td>
<td>Sialylated Complex</td>
</tr>
</tbody>
</table>

<sup>†</sup>*N*-linked glycosylation sites, excluding *N*-linked glycation.

2.3.1 Collision Induced Dissociation (CID) Studies. Representative data from glycopeptides of RNase B, asialofetuin, and transferrin, are shown in Figures 1 and 2. These data show typical fragmentation patterns for glycopeptides in the following categories: A) High mannose type, B) Complex or hybrid type and C) Complex type structures containing the more labile residues of sialic acid and/or fucose. The CID spectra of these glycopeptides illustrate that many of the same types of product ions are detected in the glycopeptide MS/MS data, regardless of the attached glycan composition. Specifically, product ions containing the peptide and portions of the pentasaccharide core are found in all these spectra and most other spectra in the training set, regardless of the glycan type. Herein, those peptide-containing product ions are
referred to as the [peptide + core component] ions.

It has been shown previously that the same [peptide + core component] ions are present in CID spectra of glycopeptides.\textsuperscript{23} From the MS/MS data we obtained, the [peptide + core component] product ions were also found to be present in multiple charge states, when the charge state of the precursor ion was greater than one, as shown in Figure 1. For all three model glycopeptides, these product ions were detected in both the precursor’s charge state and the next lowest charge states, as shown by Figure 1A – C. This finding is consistent with previous reports by Lebrilla and co-workers.\textsuperscript{27}
Figure 1. MS/MS data from model N-linked glycopeptides used to generate CID fragmentation rules with those product ions common to all appended glycans. A high mannose glycopeptide from RNase B is shown in A; a sialylated complex glycopeptide from transferrin is shown in B; and a complex glycopeptide from asialofetuin is shown in C. The spectra in A – C show the peptide-containing, or [peptide + core component], product ions detectable for all N-linked glycopeptides (regardless of the glycan attached).

The second predominant type of product ion detected in the glycopeptide data were neutral losses of terminal monosaccharides from the glycopeptide precursor ion. In contrast to the [peptide + core component] fragmentation, neutral losses from the precursor ion observed for the three model glycoprotein types were found to be unique to each candidate’s carbohydrate composition. These ions, herein referred to as the [precursor – monosaccharide] product ions, were used to develop fragmentation rules specific to glycopeptides with different glycan substituents. The fragmentation rules for both types of product ions serve as the basis for a novel algorithm and computer analysis tool that is described in Chapter 3 of this dissertation.

An example highlighting the glycan-specific fragmentation for glycopeptides is illustrated in Figure 2. In Figure 2A, a CID spectrum collected on a high mannose type glycopeptide shows sequential mannose losses, and the neutral loss of this residue as the predominant fragmentation for high mannose containing glycopeptides is well established.25, 28, 29 These ions are typically present in the spectrum in the same charge state as the precursor ion.
Figure 2. MS/MS data from model N-linked glycopeptides used to generate CID fragmentation rules with those product ions specific to the composition of an appended glycan. A high mannose glycopeptide from RNase B is shown in A; a sialylated complex glycopeptide from transferrin is shown in B; and a complex glycopeptide from asialofetuin is shown in C. The spectra in A – C show those product ions that result from neutral losses of monosaccharides, [precursor – monosaccharide], found to be unique to each N-glycan type. (Diagnostic neutral losses specific to each glycan type are shown in color, while other neutral losses that are not useful in determining the glycan type are shown in gray.)

For complex or hybrid bi- and tri-antennary structures containing no labile fucose or sialic residues, (such as the representative glycopeptide in Figure 2B), the predominant neutral losses were found to be dependent on the total number of HexNAc vs. Hex monosaccharide residues. If there are more HexNAc residues than Hex residues, the key diagnostic loss most commonly observed in the training set was shown to be loss of two HexNAc from the glycopeptide precursor ion. In comparison, those compositions containing more Hex residues than HexNAc residues showed a key diagnostic loss corresponding to the loss of [Hex + HexNAc] from the precursor ion. In addition, a glycopeptide marker ion at m/z 366 is present in the CID spectra of these compositions. Figure 2B shows an example where the [Hex + HexNAc] loss is readily detected. These characteristic fragmentation patterns were found to be essential for verifying the glycan portion of a glycopeptide.

Finally, for MS/MS data collected on glycopeptides containing labile residues such as sialic acid or fucose, the predominant [precursor – monosaccharide] product ion is the neutral loss of these labile residues from the glycopeptide precursor. For example, in Figure 2C, loss of sialic acid is detectable as a major product ion. Often, these ions are detected in both the precursor ion’s charge state, and in the charge state below that of the precursor ion. While data for only a glycopeptide containing sialic acid is shown in this chapter, glycopeptides containing at least one fucose residue generally follow the same trend, since fucose is also a more labile
monosaccharide. This idea was verified during the validation of the software program described in Chapter 3 of this dissertation, where analysis of fucosylated glycopeptide data is shown.

Although other neutral losses corresponding to the [precursor – monosaccharide] product ion types are often present in CID spectra collected on the glycopeptides, (these ions are in gray in the above figure) they were not shown to be unique enough to discriminate among various potential glycan substituent compositions.

2.3.2 N-Linked Glycopeptide Fragmentation Rules. After extensive analysis of CID spectra collected on RNase B, asialofetuin, and transferrin glycopeptides, a set of fragmentation rules to be applied for N-linked glycopeptide MS/MS data was developed. Separate fragmentation rules are implemented for the glycan portion of the glycopeptide, depending on which types of glycans are present in the candidate composition. The eight possible glycan categories include: 1) High mannose type glycans without appended fucose; 2) High mannose glycans that also contain fucose; 3) Complex or hybrid structures containing sialic acid (defined as any glycan that is not in groups 1 or 2, does not contain any fucose residues, but contains sialic acid); 4) Complex or hybrid type structures containing sialic acid and fucose residues (defined as any glycan that is not in groups 1 or 2, and contains both sialic acid and fucose residues); 5) Complex or hybrid type structures that contain fucose and multiple terminal HexNAc residues; (defined as any glycan that is not in groups 1-4, does not contain sialic acid, and has at least one fucose residue and a greater number of HexNAc than Hex residues); 6) Complex/hybrid type structures that contain fucose and terminal Hex residues (defined as any glycan that is not in groups 1-5, does not contain sialic acid, has at least one fucose residue, and has a greater number of Hex than HexNAc residues); 7) Complex/hybrid type structures with multiple terminal HexNAc residues but no sialic acid or fucose; (which is the same as group 5
glycans, except no fucose is present); and 8) Complex/hybrid type structures that lack sialic acid or fucose and contain terminal Hex residues (which is the same as group 6 glycans, except no fucose is present).

The glycan classification system described above was developed to account for the fact that glycopeptides with these different glycan components fragment differently and have different diagnostic ions identifying them, as shown herein. This approach is also supported by recently published research that shows the types of product ions in tandem mass spectra of glycopeptides vary, depending on the unique glycan substituents present. Figure 3 displays the product ions detected for each of the glycan class types devised on the basis of their fragmentation characteristics.
2.3.3 Initial Algorithm Development. After the fragmentation rules were created, as illustrated by Fig. 3, they were incorporated into a set of instructions, or algorithm, to be used to determine glycopeptide composition from a given CID spectrum. These rules were based on the detected product ions found to be present in each of the devised glycan categories. In the development of this algorithm, a variety of normalization thresholds were manually evaluated using the normalization function in the XCalibur software (Thermo-Scientific).

Normalization levels for [peptide + core component] product ions were applied by setting the relative abundance threshold to specified percentages. Extensive testing of 1%, 2%, and 3% relative abundance normalizations were performed. A relative abundance threshold of 2%
was found to work best for most spectra, however; 3% was found to be better for those spectra containing a high amount of noise whereas 1% was found to be ideal for very clean spectra with a strong signal-to-noise ratio.

For the [precursor – monosaccharide] product ions, the normalization rubric applied was more complex. The product ions formed by each of the precursor neutral losses were found to be present in varying intensities. As such, different relative abundance thresholds are applied to each unique monosaccharide loss. However, these normalization values are still based on the quality of MS/MS data and are automatically adjusted based on the peptide normalization values which are selected. Details of these normalization values are given in the original complete algorithm, as shown in Table 2.

Table 2. Original Complete Algorithm Developed from Glycopeptide Fragmentation Rules.

**INPUTS**
Spectra = MS/MS data
PrecursorIon = \( m/z \) of precursor ion
Candidate Glycan & Peptide Formulas
ChargeState = charge state of precursor ion
PeptideSearchNormalization = spectra record abundance at \{1% | 2% | 3%\}

**CONVENTIONS**
DecrementChargeState = charge state - 1
Spectra\[m/z\] = \( m/z \) value of a record in spectra
Spectra[Minima] = lowest \( m/z \) value in spectra
Spectra[Maxima] = highest \( m/z \) value in Spectra
CandidateHexNAc = number of HexNAc residues in candidate's glycan
CandidateHex = number of hexose residues in candidate's glycan
CandidateFuc = number of fucose residues in candidate's glycan
CandidateNeu5Ac = number of sialic acid residues in candidate's glycan
For all matching and scoring calculations, a match between SearchCandidate and Spectra is true only if:
(Spectra \([m/z] - 1\) \(\leq\) SearchCandidate \(\leq\) (Spectra \([m/z] + 1\))
and TotalRawPeptideScore and TotalRawGlycanScore are incremented only if:
(Spectra[Minima]) \(\leq\) SearchCandidate \(\leq\) (Spectra[Maxima])

**SPECTRA NORMALIZATION SELECTION**
A. Normalize spectra to 1%, 2%, or 3% for peptide-glycan searches:
Remove Spectra records whose relative abundance is less than or equal to PeptideSearchNormalization.
B. Normalize spectra for precursor-glycan searches:
Remove Spectra records whose relative abundance is less than or equal to the below corresponding thresholds:
If 1% PeptideSearchNormalization:
Block A1, A2: PrecursorIonSearchNormalization = 2%
Block A3, B, C, D: PrecursorIonSearchNormalization = 6%
Block E: PrecursorIonSearchNormalization = 4%
Block F: PrecursorIonSearchNormalization = 0%
If 2% PeptideSearchNormalization:
Block A1, A2: PrecursorIonSearchNormalization = 3%
Block A3, B, C, D: PrecursorIonSearchNormalization = 10%
Block E: PrecursorIonSearchNormalization = 6%
Block F: PrecursorIonSearchNormalization = 0%
If 3% PeptideSearchNormalization:
Block A1, A2: PrecursorIonSearchNormalization = 4%
Block A3, B, C, D: PrecursorIonSearchNormalization = 18%
Block E: PrecursorIonSearchNormalization = 8%
Block F: PrecursorIonSearchNormalization = 0%

1. Calculate the neutral masses of the candidates' peptides and glycans.
   PeptideMass = (masses of constituent amino acids) + 18.01056

2. Grade peptide-containing peaks.

2A. Look for each of these candidates at ChargeState:
   Y0 = (PeptideMass + (ChargeState * 1.0073)) / ChargeState
   Y1 = (PeptideMass + (ChargeState * 1.0073) + HexNAc) / ChargeState
   Y1 = (PeptideMass + (ChargeState * 1.0073) + 203.0794) / ChargeState
   Y2 = (PeptideMass + (ChargeState * 1.0073) + HexNAc + HexNAc) / ChargeState
   Y2 = (PeptideMass + (ChargeState * 1.0073) + 203.0794 + 203.0794) / ChargeState
   Y3 = (PeptideMass + (ChargeState * 1.0073) + HexNAc + HexNAc + Hex) / ChargeState
   Y3 = (PeptideMass + (ChargeState * 1.0073) + 203.0794 + 203.0794 + 162.0528) / ChargeState
   Y4 = (PeptideMass + (ChargeState * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528) / ChargeState
   Y5 = (PeptideMass + (ChargeState * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528 + 162.0528) / ChargeState
   If match, ActualRawPeptideScore += 1. Regardless of match, TotalRawPeptideScore += 1.

2B. For ChargeStateIterator from DecrementedChargeState to 1, recursively look for each of these candidates:
   Y0 = (PeptideMass + (ChargeStateIterator * 1.0073)) / ChargeStateIterator
   Y1 = (PeptideMass + (ChargeStateIterator * 1.0073) + HexNAc) / ChargeStateIterator
   Y1 = (PeptideMass + (ChargeStateIterator * 1.0073) + 203.0794) / ChargeStateIterator
   Y2 = (PeptideMass + (ChargeStateIterator * 1.0073) + HexNAc + HexNAc) / ChargeStateIterator
   Y2 = (PeptideMass + (ChargeStateIterator * 1.0073) + 203.0794 + 203.0794) / ChargeStateIterator
   Y3 = (PeptideMass + (ChargeStateIterator * 1.0073) + HexNAc + HexNAc + Hex) / ChargeStateIterator
   Y3 = (PeptideMass + (ChargeStateIterator * 1.0073) + 203.0794 + 203.0794 + 162.0528) / ChargeStateIterator
   Y4 = (PeptideMass + (ChargeStateIterator * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528) / ChargeStateIterator
   Y5 = (PeptideMass + (ChargeStateIterator * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528 + 162.0528) / ChargeStateIterator
   Y5 = (PeptideMass + (ChargeStateIterator * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528 + 162.0528) / ChargeStateIterator
   If match, ActualRawPeptideScore += 2. Regardless of match, TotalRawPeptideScore += 2.

2C. Look for the Y1 candidate at DecrementedChargeState.
   If match and relative abundance is > 25%, add 4 to ActualRawPeptideScore. Regardless of match, add 4 to TotalRawPeptideScore.

2D. PeptideScore = ActualRawPeptideScore / TotalRawPeptideScore

3. Grade precursor loss peaks.
A. Is CandidateHexNAc = 2 and CandidateHex = {1 - 9}? If yes, continue from A1. If no, continue from B.
A1. For LossMultiplier from 1 to CandidateHex, recursively look for loss of (LossMultiplier * Hex) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((LossMultiplier * 162.0528) / ChargeState)
If match, ActualRawGlycanScore += 2. Regardless of match, TotalRawGlycanScore += 2.
A2. Initialize MatchFound to true, increment LossMultiplier from its last value at A1, and while MatchFound is true, look for loss of (LossMultiplier * Hex) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((LossMultiplier * 162.0528) / ChargeState)
If match, ActualRawGlycanScore -= 2, LossMultiplier += 1.
A3. Is CandidateFuc > 0? If yes, continue from A4. If no, continue from G.
A4. Look for loss of (Fuc) from PrecursorIon at DecrementChargeState:
SearchCandidate = ((PrecursorIon * ChargeState) - 1.0073 - 146.0579) / DecrementChargeState
If match, ActualRawGlycanScore += 2. Regardless of match, TotalRawGlycanScore += 2.
A5. For LossMultiplier from 1 to CandidateFuc, recursively look for loss of (LossMultiplier * Fuc) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((LossMultiplier * 146.0579) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
A6. Continue from G.

B. Is CandidateNeu5Ac > 0 and CandidateFuc > 0? If yes, continue from B1. If no, continue from C.
B1. Look for loss of (Neu5Ac + Fuc) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((291.0954 + 146.0579) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
B2. Continue from C.

C. Is CandidateNeu5Ac > 0? If yes, continue from C1. If no, continue from D.
C1. Look for loss of (Neu5Ac) from PrecursorIon at DecrementChargeState:
SearchCandidate = (PrecursorIon * ChargeState) - 1.0073 - 291.0954) / DecrementChargeState
If match, ActualRawGlycanScore += 2. Regardless of match, TotalRawGlycanScore += 2.
C2. For LossMultiplier from 1 to CandidateNeu5Ac, recursively look for loss of (LossMultiplier * Neu5Ac) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((LossMultiplier * 291.0954) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
C3. Continue from D.

D. Is CandidateFuc > 0? If yes, continue from D-1A. If no, continue from D-2.
D-1A. Look for loss of (Fuc) from PrecursorIon at DecrementChargeState:
SearchCandidate = ((PrecursorIon * ChargeState) - 1.0073 - 146.0579) / DecrementChargeState
If match, ActualRawGlycanScore += 2. Regardless of match, TotalRawGlycanScore += 2.
D-1B. For LossMultiplier from 1 to CandidateFuc, recursively look for loss of (LossMultiplier * Fuc) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((LossMultiplier * 162.0528) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
D-1C. Is CandidateNeu5Ac > 0? If no, continue from D-1D. If yes, continue from G.
D-1D. Subtract pentasaccharide core ([HexNAc]2Hex[3]) from glycan. Of the remaining sugars, are there at least 2 more HexNAc than Hex residues? If yes, continue from D-1D-1A. If no, continue from D-1D-2A.
D-1D-1A. Look for loss of (HexNAc) from PrecursorIon at DecrementChargeState:
SearchCandidate = ((PrecursorIon * ChargeState) - 1.0073 - 146.0579 - 203.0794) / DecrementChargeState
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
D-1D-1B. Look for loss of (Fuc + HexNAc) from PrecursorIon at ChargeState:
SearchCandidate = (PrecursorIon * ChargeState) - 1.0073 - 146.0579 - 203.0794) / ChargeState
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
D-1D-1C. Is CandidateFuc > 1? If yes, continue from D-1D-1D. If no, continue from D-1D-1E.
D-1D-1D. Look for loss of ((CandidateFuc * Fuc) + HexNAc) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - (((CandidateFuc * 146.0579) + 203.0794) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
D-1D-1E. Look for loss of ((CandidateFuc * Fuc) + (2 * HexNAc)) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - (((CandidateFuc * 146.0579) + (2 * 203.0794)) / ChargeState)
If match, ActualRawGlycanScore += 2. Regardless of match, TotalRawGlycanScore += 2.
D-1D-1F. Continue from G.
D-1D-2A. Look for loss of (Fuc + Hex) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((146.0579 + 162.0528) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
D-1D-2B. Is CandidateFuc > 1? If yes, continue from D-1D-2C. If no, continue from D-1D-2D.
D-1D-2C. Look for loss of (((CandidateFuc * Fuc) + Hex) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - (((CandidateFuc * 146.0579) + 162.0528) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
D-1D-2D. Look for loss of (Fuc + Hex + HexNAc) from PrecursorIon at DecrementedChargeState:
SearchCandidate = ((PrecursorIon * ChargeState) - 1.0073 - 146.0579 - 162.0528 - 203.0794) / DecrementedChargeState
If match, continue from D-1D-2D-1A. If no match, continue from D-1D-2D-2A.
D-1D-2D-1A. Add 4 to ActualRawGlycanScore. Add 4 to TotalRawGlycanScore.
D-1D-2D-1B. Continue from F1.
D-1D-2D-2A. Look for loss of (Fuc + Hex + HexNAc) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((146.0579 + 162.0528 + 203.0794) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
D-1D-2D-2B. Continue from F1.
D-2. Is CandidateNeu5Ac > 0? If no, continue from E. If yes, continue from G.
E. Subtract pentasaccharide core ([HexNAc]2Hex[3]) from glycan. Of the remaining sugars, are there at least 2
more HexNAc than Hex residues? If yes, continue from E-1A. If no, continue from E-2A.
E-1A. Look for loss of (HexNAc) from PrecursorIon at DecrementedChargeState:
SearchCandidate = ((PrecursorIon * ChargeState) - 1.0073 - 203.0794) / DecrementedChargeState
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
E-1B. Look for loss of (HexNAc) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - (203.0794 / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
E-1C. Look for loss of (2 * HexNAc) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((2 * 203.0794) / ChargeState)
If match, ActualRawGlycanScore += 2. Regardless of match, TotalRawGlycanScore += 2.
E-1D. Continue from G.
E-2A. Look for loss of (Hex) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - (162.0528 / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
E-2B. Look for loss of (Hex + HexNAc) from PrecursorIon at DecrementedChargeState:
SearchCandidate = ((PrecursorIon * ChargeState) - 1.0073 - 162.0528 - 203.0794) / DecrementedChargeState
If match, continue from E-2B-1A. If no match, continue from E-2B-2A.
E-2B-1A. Add 4 to ActualRawGlycanScore. Add 4 to TotalRawGlycanScore.
E-2B-1B. Continue from F1.
E-2B-2A. Look for loss of (Hex + HexNAc) from PrecursorIon at ChargeState:
SearchCandidate = PrecursorIon - ((162.0528 + 203.0794) / ChargeState)
If match, ActualRawGlycanScore += 4. Regardless of match, TotalRawGlycanScore += 4.
E-2B-2B. Continue from F1.
F1. Look for Hex-HexNAc marker ion (366).
If match, ActualRawGlycanScore += 2. Regardless of match, TotalRawGlycanScore += 2.
3F2. Go to G.
3G. GlycanScore = ActualRawGlycanScore / TotalRawGlycanScore
4. GlycopeptideScore = (PeptideScore * 0.67) + (GlycanScore * 0.33)
2.4 CONCLUDING REMARKS

Numerous CID experiments on glycopeptides containing various N-linked glycan types were performed. Two main types of product ions were subsequently identified from the resultant of MS/MS data, one of which was found to be present in all of the glycopeptide data studied, and one of which was found to be unique to the arrangement of the attached glycan. These product ions are referred to as [peptide + core component] ions and [precursor – monosaccharide] ions, respectively. After studying the collection of spectra, a set of fragmentation rules to be applied to each of the eight devised glycan type categories was developed. These fragmentation rules were then incorporated into an algorithm that functions to predict two types of product ions from CID data of varying spectral quality. The algorithm was eventually turned into a publicly available automated analysis tool, which is described in Chapter 3 of this dissertation.

2.5 ACKNOWLEDGEMENTS

The author acknowledges financial support from the NIH (RO1RR026061) to H.D., an NSF Fellowship (DGE-0742523) to C.W. and K.R., and a Pfizer Award to C.W.

The author also wishes to thank those who contributed to the work described herein: David Hua for writing the final version of the algorithm, Morgan Maxon for all of her work and time spent interpreting data, Katie Rebecchi for collecting and providing some of the N-linked glycopeptide CID spectra analyzed, and Heather Desaire for her guidance and dedication.
2.6 REFERENCES


(28) Alley, W. R.; Mechref, Y.; Novotny, M. V. Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. *Rapid

CHAPTER 3

GLYCOPEP GRADER: A WEB-BASED UTILITY FOR ASSIGNING THE COMPOSITION OF N-LINKED GLYCOPEPTIDES

The work described in Chapter 2 encompasses an original (first author) publication:


ABSTRACT

GlycoPep Grader (GPG) is a freely available software tool designed to accelerate the process of accurately determining glycopeptide composition from tandem mass spectrometric data. GPG relies on the identification of unique dissociation patterns shown for high mannose, hybrid, and complex N-linked glycoprotein types, including patterns specific to those structures containing fucose or sialic acid residues. The novel GPG scoring algorithm scores potential candidate compositions of the same nominal mass against MS/MS data through evaluation of the Y₁ ion and other peptide-containing product ions, across multiple charge states, when applicable. In addition to evaluating the peptide portions of a given glycopeptide, the GPG algorithm predicts and scores product ions that result from unique neutral losses of terminal glycans. GPG has been applied to a variety of glycoproteins, including RNase B, asialofetuin and transferrin, and the HIV envelope glycoprotein, CON-S gp140 CFI. The GPG software is implemented predominantly in PostgreSQL, with PHP as the presentation tier, and is publically accessible online. Thus far, the algorithm has identified the correct compositional assignment from multiple candidate N-glycopeptides in all tests performed.
3.1 INTRODUCTION

Among all co/post-translational modifications, glycosylation is widely regarded as both the most frequent and most complex that proteins undertake.\textsuperscript{1, 2, 3, 4} It is well-documented that glycosylation regulates a variety of intra- and extra-cellular processes.\textsuperscript{3, 4, 5, 6, 7, 8, 9} Cellular communication and transport events,\textsuperscript{5, 6} and mechanisms of protein folding,\textsuperscript{3, 5, 6} degradation,\textsuperscript{3, 5} and enzymatic interaction,\textsuperscript{7} have all been shown to be regulated by glycosylation, the majority of which are $N$-linked in type.\textsuperscript{1} As such, the availability of mass spectrometry (MS) tools to speed the identification of glycosylation profiles is critical to the elucidation of their physiological importance.\textsuperscript{3, 8, 10, 11, 12}

Typically, glycosylation analysis using mass spectrometry (MS) techniques is accomplished using one of two approaches: Glycan analysis and glycopeptide analysis.\textsuperscript{12} The most information-rich of these methods is glycopeptide analysis, as glycosylation characteristics at individual sites of glycan attachment are readily identifiable.\textsuperscript{2, 12} High resolution MS data is used to determine potential candidate compositions for mass spectral peaks that are suspected or known to be from glycopeptides. Computer-based programs such as Glycomod\textsuperscript{13} and GlycoPepDB\textsuperscript{14} calculate glycopeptide candidate compositions on the basis of mass information, as do a number of custom-generated databases.\textsuperscript{15, 16} Unfortunately, a large amount of mass redundancy is typically encountered in glycopeptide analysis. Many different combinations of glycan composition + peptide composition are isobaric,\textsuperscript{15} so multiple candidate compositions frequently correspond to the same nominal mass. Therefore, while high resolution MS data is useful for predicting possible glycopeptide candidate compositions, it alone is not sufficient to identify glycopeptides unambiguously. As a result, MS/MS experiments are often necessary to correctly assign glycopeptide compositions. When the analyses of these data are performed manually, the
A few unique strategies have been developed to automate the process of scoring MS/MS data against potential glycopeptide compositions. These include programs described in references 18, 19, 20, 21, 22. However, none of these analysis tools are freely accessible to the public. In terms of those tools that are publicly available for glycopeptide analysis, many have been designed to predominantly analyze the fragmentation of glycans. Although these tools are capable of analyzing glycopeptides, the peptide component must be known in advance, which severely limits their utility for analysis of unknown glycopeptides.

GlycoWorkBench and Glyco-Peakfinder both utilize this approach for the annotation of glycans in glycopeptide data. A completely different approach is utilized by GlycoPep ID, a web-based tool developed by Go et al. GlycoPep ID interprets MS/MS data of glycopeptides to identify the peptide component of glycopeptides through analysis of expected product ions, but the key disadvantage of this program is that it does not include a scoring function.

The most promising publicly accessible tool specifically developed to interpret and score MS/MS data of glycopeptides is GlycoMiner, developed by Ozohanics et al. This program was designed to analyze qTOF data, and is capable of identifying and assigning glycopeptide compositions when both the peptide and glycan portions are unknown. Although this program is a great advancement in the automation of glycopeptide MS/MS analysis, GlycoMiner often generates multiple plausible compositions and fails to rank the correct glycopeptide as the top candidate, instead listing it as one of the most probable compositions. In addition, the program requires the presence of low-mass marker ions, which are generally not present in data collected on ion trap instruments. The program also requires the MS/MS data to be transformed into singly charged ions, prior to analysis. This transformation is often not possible when analyzing
low resolution MS/MS data, such as that from an ion trap mass spectrometer. Finally, GlycoMiner requires MS/MS data containing a low S/N.\textsuperscript{26}

GlycoPep Grader, which aims to expedite the characterization of N-linked glycopeptides by evaluating both the glycan and peptide portions through a series of devised fragmentation rules, was developed in an effort to overcome the limitations of the currently available tools. The novel algorithm calculates and scores any given glycopeptide candidate composition by searching MS/MS data for two types of product ions: 1) Those containing the peptide portion, [peptide + core component] ions, and 2) Those resulting from neutral losses of terminal monosaccharides, [precursor – monosaccharide] ions. The use of GlycoPep Grader in determining glycopeptide compositions is not contingent upon any spectral requirements, such the presence of specific marker ions. In addition, the GPG algorithm analyzes MS/MS data in a charge-state dependent fashion, bypassing the need for transformation of spectra to singly-charged ions. These features have resulted in a highly accurate automated analysis tool that deciphers glycopeptide compositions. GPG is freely available online; it can be accessed at http://glycopro.chem.ku.edu/GPGHome.php.

3.2 EXPERIMENTAL

3.2.1 Materials and Reagents. Details regarding the materials and reagents, along with the experimental protocols for sample preparation and MS analysis of RNase B, asialofetuin, and transferrin glycopeptides can be found in the experimental section of Chapter 2 of this dissertation.

3.2.2 Production of CON-S gp140 CFI Glycoprotein. CON-S gp140 CFI envelope glycoprotein was obtained by our lab from the Duke Human Vaccine Research Institute (Durham, NC) after it was constructed, expressed and purified using methods previously
3.2.3 Preparation and LC-MS of CON-S gp140 CFI Glycopeptides. Purified envelope glycoprotein samples were prepared by Go et al. as stated in the literature. Briefly, 300 μg aliquots of glycoprotein were denatured by the addition of 6 M urea in 100 mM Tris buffer (pH 7.5) with 3 mM EDTA. The denatured proteins were then reduced and alkylated by incubation in 15 mM DTT at room temperature for 1 hr. Immediately after, 40 mM IAM was allowed to react with each denatured sample at room temperature in the dark for an additional 1 hr. To neutralize excess IAM, a second portion of DTT was added to achieve a final concentration of 50 mM DTT. After reduction and alkylation, the samples were diluted to 2 M urea prior to adding trypsin at a protein/enzyme ratio of 30:1 (w/w). The protease was allowed to react at 37 °C overnight, followed by a second trypsin digestion under the same conditions. HIV Env glycopeptides were then subjected to LC-MS and identified by Go et al., as described. Finally, the resultant collection of CID spectra was scored using GPG.

3.2.4 Development of a Glycopeptide Training Data Set. In order to develop the GPG algorithm, a set of “known” glycopeptides and their MS/MS data were required; the training set included glycopeptides from RNase B, asialofetuin and transferrin, as these are well characterized samples. To identify the glycopeptides from these samples in the MS data, a prediction table of theoretical m/z values corresponding to glycopeptide compositions for each of the three proteins was prepared. The amino acid sequences from RNase B, asialofetuin, and transferrin were obtained from Uniprot (www.uniprot.org) and their sequences were imported into Protein Prospector (http://prospector.ucsf.edu/prospector/mshome.htm) where tryptic peptides containing Cys residues were modified with carbamidomethylation, and a theoretical tryptic digest was performed to consider up to two tryptic miscleavages. The masses of the
peptides that contained potential $N$-linked glycosylation sites were added to the masses of the known glycan compositions for each glycosylation site, in order to obtain glycopeptide masses. These masses were converted into $m/z$ values corresponding to the glycopeptides in multiple charge states. The MS/MS data for RNase B, asialofetuin, and transferrin were then searched to identify spectra that corresponded to the correct $m/z$ value for a given glycopeptide composition. The MS/MS data were carefully (manually) evaluated, to verify the glycopeptide assignment.

3.2.5 The Glycopeptide Validation Data Set. In order to test the GPG software, a validation set of glycopeptide compositions that were not used in the fragmentation studies or algorithm development was necessary. The validation set for these studies comprised data from a glycoprotein, CON-S gp140ΔCFI, which had been previously analyzed in our laboratory. Data from this protein was selected because prior analyses demonstrated that all the necessary glycoform types were present as glycopeptides (including high mannose and complex/hybrid structures with and without sialic acid and fucose.) Additionally, since the protein has more than 25 glycosylation sites, a wide variety of glycosylated peptide sequences were also available. Furthermore, all the MS/MS data on this protein had been previously analyzed manually, as described elsewhere.

3.2.6 Software Platform. GlycoPep Grader is a Web service implementation of our algorithm, encapsulating data submission and analysis as a computational session. This transaction-processing approach protects our Web service against the thankless perils that come with providing anonymous data acceptance and computational services on the Internet, while simultaneously ensuring the correctness of the computation. The graphical user interface (GUI) code is built to conform to ECMAScript and W3C DOM standards, and we chose the open-source, globally distributed Mozilla Firefox Web browser as the reference platform for the GUI
presentation. The computational engine is implemented on common Web server and database software, with a variety of implementation-specific optimizations for computationally-intensive hotspots in the algorithm. These optimizations include deep logic reordering, pre-calculations of elicited constants, and pre-compilations of common loops. Finally, we use AJAX technology (Asynchronous Javascript And XML) to achieve state continuity and provide a responsive, interactive experience to the user.

Prior to using GlycoPep Grader, the user must first successfully complete a simple math problem embedded in a CAPTCHA (completely automated public Turing test to tell computers and humans apart). This security step helps prevent automated abuse of the Web server. GlycoPep Grader then accepts user input, including candidate glycopeptide compositions, the \( m/z \) and charge state of the precursor ion, and MS/MS data (which the user provides in a .CSV file). The Web service performs server-side validation of the submitted data for type, format, size, and range correctness. Once the data obtains correctness approval, the computational engine performs its analysis of the glycopeptide candidates against the spectral data. When the analysis is complete, the computational engine assembles and returns the results to the GUI code listening on the user's Firefox Web browser.

3.2.7 Generation and Input of Glycopeptide Candidate Compositions. After the MS/MS peak list file (along with the corresponding charge state and \( m/z \) of the precursor ion) is uploaded to GPG in .CSV file format, peptide compositions are input manually by listing the amino acid sequence of each glycopeptide candidate ion vertically on a separate line. The glycopeptide candidate compositions are obtained by the user through freely accessible programs such as GlycoMod or GlycoPep DB, or custom-generated databases. The GPG analysis tool then quickly calculates and searches for the [peptide + core component] product ions that it
predicts to be present for each of the peptide portions entered. In the next window, the glycan portions for each of the candidate glycopeptides are manually entered in the same order using the following format, where \( n \) = the number of each monosaccharide residue and Neu5Ac = sialic acid: \([\text{HexNAc}]_n[\text{Hex}]_n[\text{Neu5Ac}]_n[\text{Fuc}]_n\). After GPG evaluates the uploaded MS/MS peak list for product ions expected to be present for each glycan, a final score is displayed in the output for each of the user-entered glycopeptide compositions.

### 3.2.8 False Discovery Rate Determination and Scoring of Candidate Compositions.

Decoy candidate compositions for all data sets were generated using an in-house database where a decoy polypeptide of 50,000 amino acid residues, *Titin*, was multiplexed to a biologically relevant library of approximately 200 glycans. (These glycans are the same ones used in the online tool, GlycoPep DB.) All selected decoy candidate compositions have a calculated neutral mass that is within 50 ppm of the FT-ICR MS monoisotopic peak value of the glycopeptide precursor ion for the CID spectrum tested. The decoy glycopeptide compositions, along with the correct glycopeptide composition assignment, were used to determine the false discovery rate of the GPG tool.

### 3.3 RESULTS AND DISCUSSION

GlycoPep Grader (GPG) was designed to analyze N-linked glycopeptide CID data. RNase B, asialofetuin, and transferrin were chosen as model glycoproteins for the initial testing of the GPG software tool, as well as for the development of the novel algorithm that powers GPG, because they are well characterized and contain various glycoform types. Detailed information on the glycosylation characteristics of the glycopeptides used for the testing and validation of the GPG software tool, is included below in Table 1.
Table 1. Glycopeptides Analyzed for Training and Validation Data Sets.

<table>
<thead>
<tr>
<th>Glycoprotein</th>
<th>Mass (Da)</th>
<th>Length (AA)</th>
<th># N-glycan Sites</th>
<th>N-Glycan Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNase B&lt;sup&gt;1&lt;/sup&gt;</td>
<td>16,461</td>
<td>150</td>
<td>1</td>
<td>High Mannose</td>
</tr>
<tr>
<td>Asialofetuin&lt;sup&gt;1&lt;/sup&gt;</td>
<td>38,419</td>
<td>359</td>
<td>3</td>
<td>Complex</td>
</tr>
<tr>
<td>Transferrin&lt;sup&gt;1&lt;/sup&gt;</td>
<td>77,064</td>
<td>698</td>
<td>2</td>
<td>Sialylated Complex</td>
</tr>
<tr>
<td>CON-S gp140ΔCFI&lt;sup&gt;2&lt;/sup&gt;</td>
<td>~140,000</td>
<td>610</td>
<td>21</td>
<td>Highly Diverse</td>
</tr>
</tbody>
</table>

<sup>1</sup> Training Data Set  
<sup>2</sup> Validation Data Set

### 3.3.1 Novel GPG Scoring Algorithm

A detailed version of the scoring system, as it stands currently, is available at http://glycopro.chem.ku.edu/GPGHome.php. The original scoring algorithm is also available in Chapter 2 of this dissertation. The same peptide-containing product ions are detected in a CID spectrum of an N-linked glycopeptide, regardless of the type of glycan substituent attached. Therefore, for GlycoPep Grader (GPG) scoring of each peptide portion, the [peptide + core component] product ions are calculated for the candidate glycopeptide beginning with the [naked peptide] and continuing through the [peptide + intact pentasaccharide core] for a total six possible [peptide + core component] product ions: 1. [naked peptide], 2. [peptide + HexNAc], 3. [peptide + 2HexNAc], 4. [peptide + 2HexNAc + Hex], 5. [peptide + 2HexNAc + 2Hex], and 6. [peptide + 2HexNAc + 3Hex]. The GPG algorithm uses the presence of these ions to score the peptide portion of the candidate glycopeptide composition. The Y<sub>1</sub> ion, which contains the peptide and one HexNAc residue from the pentasaccharide core, has been shown to be a highly abundant ion in MS/MS data collected on glycopeptides. This product ion is also considered a very indicative identifier of a glycopeptide’s peptide portion, so the GPG algorithm weights this ion more heavily and scores it on the basis of its intensity as well. Each of these ions is then searched for in the MS/MS data in multiple charge states. The scoring algorithm for these ions does not change, regardless of the N-linked glycopeptide type.
A separate GPG scoring scheme is implemented for the glycan portion of a glycopeptide, depending on which type of glycan is present in each candidate composition. These eight glycan categories and the diagnostic product ions expected to be detected for each are described in Chapter 2 of this dissertation as well.

In addition to determining which diagnostic ions should be scored for each of the candidate glycopeptide compositions, we have implemented noise-reduction and intensity-based scoring components into the algorithm. A baseline noise correction is applied before the automatic “spectral match searching” is performed in order to limit false positive peak matches arising from noise. In preliminary testing, a cut-off of 2 % has been found to be ideal for most spectra, but the algorithm allows the user to vary this cut-off, so that spectra of differing quality (noise levels) can be scored using different thresholds for noise reduction.

The relative abundance of the [precursor – monosaccharide] product ions is also taken into account when determining whether or not a peak corresponding to a particular m/z is actually from the neutral loss being evaluated, with varying threshold limits being applied according to the composition of the monosaccharide resides in the neutral loss being scored. For example, as fucose and sialic acid are more labile than Hex or HexNAc residues, the threshold applied to the detection of product ions resulting from cleavage of these residues is much higher than the threshold applied to the scoring of product ions that arise from the cleavage of Hex and HexNAc residues. This feature was implemented to reduce the possibility of false positive matches. Detailed information on the normalization thresholds used in the scoring scheme can be found in the complete algorithm, located in Chapter 2 of this dissertation.

3.3.2 Candidate Composition Scoring by GPG. After algorithm development, MS/MS data of glycopeptide spectra from RNase B, asialofetuin, and transferrin were scored using the
GPG software. The resultant collection of CID spectra obtained during the fragmentation studies described in Chapter 2 of this dissertation is referred to herein as the training data set. For each case, the known composition of the glycopeptide was scored against at least three decoy compositions, which were generated as described in the experimental section. Glycopeptide data from a variety of precursor charge states were scored.

In Figure 1, an example of the candidate composition scoring by GPG is shown for a CID spectrum collected from a high mannose type glycopeptide from RNase B. The same spectrum is shown in Figure 1A, B, and C. However, each panel shows a different candidate composition for this spectrum and includes the results of how GPG scored each composition. The correct composition is in 1A, while two decoy compositions are shown in Figure 1B and 1C. The [precursor – monosaccharide] product ions searched by GPG are calculated based on the candidate composition. For candidate A, which contains a high mannose glycan, GPG predicts the sequential loss of mannose residues from the precursor ion and evaluates the [precursor – monosaccharide] product ions by searching the MS² peak list for the m/z values corresponding to sequential losses of individual hexose residues. Candidate compositions B and C are both classified as complex or hybrid glycans without sialic acid or fucose, so the same set of fragmentation rules applies for the glycan component in these two spectra. In addition to variations in the glycan scoring, each spectrum is scored differently for the [peptide + core component] ions, because each spectrum has a different candidate peptide composition. As a result, GPG returns separate scores for the candidate compositions in B and C, even though the glycan portions are similar. The calculations for the different types of fragmentation ions are weighted by the software, with [peptide + core component] product ions accounting for 67 % and [precursor – monosaccharide] product ions accounting for 33 % of the score. GPG reports a
final score of 97 % for the correct glycopeptide assignment (candidate composition A), 20 % for the first decoy glycopeptide assignment (candidate composition B) and 27 % for the second decoy glycopeptide assignment (candidate composition C).
Figure 1. CID data of an RNase B glycopeptide from the training data set. (A) GPG scoring of the correct glycopeptide composition: 97%. (B), (C) Scoring of two decoy compositions of the same nominal mass: 20% and 27%, respectively. Exact neutral masses of candidate compositions shown in (A), (B), and (C) are, in order: 2419.9945, 2419.9733, and 2419.9978. The X on arrows in spectra indicate the absence of a product ion that was predicted to be present by GPG for a given candidate composition. A relative abundance threshold of 2% was used for [peptide + core component] product ion matching to decrease false positives from noise.

A second example of spectra scored by GPG is presented in Figure 2. The MS/MS data shown here is a sialylated glycopeptide from transferrin. In this case, different types of [precursor – monosaccharide] product ions are searched, as different glycans are present in the candidate and decoy compositions. GPG scoring of the actual transferrin glycopeptide, on which the CID spectrum was obtained, is shown in Figure 2A. In Figure 2B and 2C, the same spectrum is shown, along with two decoy glycopeptide compositions that have the same nominal mass as the correct composition. To score each spectrum, the [precursor – monosaccharide] product ions searched by GPG are determined based on each candidate’s composition, as depicted in Figure 2.

For candidate A, a sialylated complex type glycopeptide that does not contain fucose, the GPG algorithm searches for the loss of a sialic acid residue in both the charge state of the precursor and in the charge state below the precursor. As the correct composition contains two sialic residues (candidate A), GPG also searches for a loss of two sialic acid residues in the precursor charge state.

The decoy composition in Figure 2B is classified as a complex/hybrid composition that lacks sialic acid or fucose and contains more Hex than HexNAc residues. Therefore, the GPG software evaluates the presence or absence of the same [precursor – monosaccharide] product ions that were described for candidates B and C in Fig. 1. Likewise, as the decoy composition in Figure 2C is classified as a high mannose type, the [precursor – monosaccharide] product ions that GPG evaluates are neutral losses of hexose residues, as detailed in Fig. 1 for candidate
composition A. In comparison to the correct glycopeptide composition (candidate A), a lower number expected ions evaluated by GPG are found in the MS/MS data for candidates B and C, resulting in a lower GPG score.
Figure 2. CID spectrum of a sialylated transferrin glycopeptide from the training set and GlycoPep Grader’s scoring of candidate glycopeptides for m/z 1841. (A) Shows the scoring of the “correct” glycopeptide composition, while (B) and (C) show the GPG scoring mechanism applied to two decoy candidates of the same nominal mass. Arrows marked with X indicate ions that were not present in the peak list for the spectrum shown. GPG assigned a score of 75 % to the actual glycopeptide composition of CGLVPVLAENYNK + [HexNAc]4[Hex]5[Neu5Ac]2, candidate A. GPG returned a score of score 7 % and 39 % to the two decoy compositions of ETTRVNVSSSK + [HexNAc]5[Hex]9 and NLTEGEEYTFQVMANVSAGRSAPR + [HexNAc]2[Hex]4, candidates B and C, respectively. Thus, GPG analysis determined the correct glycopeptide candidate to be the most probable glycopeptide composition based on presence or absence of calculated CID product ions expected to be present for each glycopeptide scored. A threshold of 2 % relative abundance was used as the cut-off for a matching ion detected in the MS² data peak list.

In Table 2, 45 test examples are provided that show GPG scores for glycopeptides analyzed from experimental MS/MS data in the training data set. For each example, the correct composition is compared against at least two decoy compositions of the same nominal mass. A wide variety of glycopeptide compositional arrangements were tested. Over 150 glycopeptide spectra from the 45 unique glycopeptides in the training data set were scored using GPG, with the correct candidate receiving the highest score in each test performed.
Table 2. Score Results Calculated by GPG Software for Tests Performed on CID Spectra in the Training Data Set.

<table>
<thead>
<tr>
<th>Test</th>
<th>Charge</th>
<th>Candidate</th>
<th>m/z</th>
<th>Glycopeptide Composition</th>
<th>GPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2+</td>
<td>A</td>
<td>967.9252</td>
<td>SRNLTK + [HexNAc]2[Hex]5</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>967.8831</td>
<td>NASHK + [HexNAc]2[Hex]6</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>967.9336</td>
<td>WVRHNK + [HexNAc]3[Hex]3</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>2+</td>
<td>A</td>
<td>981.9227</td>
<td>NLTKDR + [HexNAc]2[Hex]5</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>981.9125</td>
<td>NRSLTN + [HexNAc]3[Hex]4</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>981.9550</td>
<td>RETQAVNWTK + [HexNAc]2[Hex]2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2+</td>
<td>A</td>
<td>1048.9516</td>
<td>SRNLTK + [HexNAc]2[Hex]6</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1048.9095</td>
<td>NASHK + [HexNAc]2[Hex]7</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1048.9600</td>
<td>WVRHNK + [HexNAc]3[Hex]4</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>D</td>
<td>1048.9492</td>
<td>NVTGTTSETIK + [HexNAc]3[Hex]3</td>
<td>44</td>
</tr>
<tr>
<td>4</td>
<td>2+</td>
<td>A</td>
<td>1103.4892</td>
<td>SRNLTKDR + [HexNAc]2[Hex]5</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1103.6655</td>
<td>NRSLTN + [HexNAc]5[Hex]3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1103.4863</td>
<td>PFNQTKNRF + [HexNAc]2[Hex]4</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>2+</td>
<td>A</td>
<td>1129.9780</td>
<td>SRNLTK + [HexNAc]2[Hex]7</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1129.9756</td>
<td>NVDSVNVGTCR + [HexNAc]3[Hex]3</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1129.9674</td>
<td>AYANVSSK + [HexNAc]3[Hex]5</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>D</td>
<td>1129.9864</td>
<td>WVRHNK + [HexNAc]3[Hex]5</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>2+</td>
<td>A</td>
<td>1143.9755</td>
<td>NLTKDR + [HexNAc]2[Hex]7</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1144.0078</td>
<td>RETQAVNWTK + [HexNAc]2[Hex]4</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1144.0129</td>
<td>NVDSVNVGTCR + [HexNAc]3[Hex]3</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>D</td>
<td>1143.9516</td>
<td>ANK + [HexNAc]5[Hex]4[Neu5Ac]1</td>
<td>27</td>
</tr>
<tr>
<td>7</td>
<td>2+</td>
<td>A</td>
<td>1184.5156</td>
<td>SRNLTKDR + [HexNAc]2[Hex]6</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1184.5127</td>
<td>PFNQTKNRF + [HexNAc]2[Hex]5</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1184.5015</td>
<td>VNVSSSSK + [HexNAc]5[Hex]3[Neu5Ac]1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2+</td>
<td>A</td>
<td>1211.0044</td>
<td>SRNLTK + [HexNAc]2[Hex]8</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1211.0128</td>
<td>WVRHNK + [HexNAc]3[Hex]6</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1211.0020</td>
<td>NVDSVNVGTCR + [HexNAc]3[Hex]3</td>
<td>33</td>
</tr>
<tr>
<td>9</td>
<td>2+</td>
<td>A</td>
<td>1225.0019</td>
<td>NLTKDR + [HexNAc]2[Hex]8</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1225.0207</td>
<td>NLTKDRGP + [HexNAc]3[Hex]4[Neu5Ac]1</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1225.0393</td>
<td>NVTGTTSETIK + [HexNAc]4[Hex]3</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>D</td>
<td>1225.0342</td>
<td>RETQAVNWTK + [HexNAc]2[Hex]5</td>
<td>47</td>
</tr>
<tr>
<td>10</td>
<td>2+</td>
<td>A</td>
<td>1265.5420</td>
<td>SRNLTKDR + [HexNAc]2[Hex]7</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1265.5183</td>
<td>NRSLTN + [HexNAc]5[Hex]5</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1265.5309</td>
<td>VNVSSSSK + [HexNAc]5[Hex]4[Neu5Ac]1</td>
<td>25</td>
</tr>
<tr>
<td>11</td>
<td>2+</td>
<td>A</td>
<td>1292.0308</td>
<td>SRNLTK + [HexNAc]2[Hex]9</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1292.0284</td>
<td>NVDSVNVGTCR + [HexNAc]3[Hex]5</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1292.0392</td>
<td>ANK + [HexNAc]5[Hex]4[Neu5Ac]1</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>2+</td>
<td>A</td>
<td>1306.0283</td>
<td>NLTKDR + [HexNAc]2[Hex]9</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>B</td>
<td>1306.0654</td>
<td>NLTKDR + [HexNAc]4[Hex]4[Neu5Ac]1</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>C</td>
<td>1306.0781</td>
<td>LVINR + [HexNAc]6[Hex]3[Neu5Ac]1</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>2+</td>
<td>A</td>
<td>1346.5684</td>
<td>SRNLTKDR + [HexNAc]2[Hex]8</td>
<td>76</td>
</tr>
</tbody>
</table>
2 + B 1346.5447 NRSLTN + [HexNAc]5[Hex]6
2 + C 1346.6205 LNVTLWK + [HexNAc]4[Hex]3[Neu5Ac]1
2 + D 1346.5913 DNGSPILGYWLEK + [HexNAc]2[Hex]4[Fuc]1
14 + A 1059.6808 RPTGEVYDIEIDTLETTCHVLDPDPLANSVR + [HexNAc]4[Hex]5
5 + B 1059.6980 IENTTTVLKSSATFQSTVAGSSPISITWLK + [HexNAc]5[Hex]5[Neu5Ac]1
5 + C 1059.6567 CHYMTIHNVTPDEGVYSVIARLEPR + [HexNAc]5[Hex]6[Neu5Ac]1
5 + D 1059.6535 NAAGNFESPSGDSSGATARDEIDAPNALSEDPK + [HexNAc]4[Hex]6[Fuc]2
15 + A 1132.7073 RPTGEVYDIEIDTLETTCHVLDPDPLANSVR + [HexNAc]5[Hex]6
5 + B 1132.6652 SCEPVPARDPCDPPQPEVTNITR + [HexNAc]6[Hex]7[Neu5Ac]2[Fuc]1
5 + C 1132.7444 IENTTTVLKSSATFQSTVAGSSPISITWLK + [HexNAc]6[Hex]6[Neu5Ac]1
5 + D 1132.6708 HILVINDQFDEGVYTAEVEGK + [HexNAc]6[Hex]7[Neu5Ac]1
16 + A 1160.5442 NVTVIEGESVTLECHISGYPSPTVTWYR + [HexNAc]5[Hex]5
4 + B 1160.5187 TDTMRLLERPPEFTLPLYNK + [HexNAc]4[Hex]5[Neu5Ac]2
4 + C 1160.5254 NTVIEGESVTLECHISGYPSPTVTWYR + [HexNAc]5[Hex]3
4 + D 1160.5186 LTPESTREFLCINGSHIQPLK + [HexNAc]4[Hex]8
17 + A 1243.5312 KLCPDCLPLAPLNDSP + [HexNAc]5[Hex]6
3 + B 1243.5153 YDSGKYTLTLENSSGTK + [HexNAc]4[Hex]8
3 + D 1243.5069 EQAVNWTK + [HexNAc]4[Hex]5[Neu5Ac]3
19 + A 1286.2296 KLCPDCLPLAPLNDSP + [HexNAc]5[Hex]6
3 + B 1286.2375 EFLCINGSHIQPLK + [HexNAc]8[Hex]3
3 + C 1286.1893 DSVNLTWTEPSDGGSK + [HexNAc]4[Hex]7[Fuc]1
3 + D 1286.1844 KAYATITNNCTK + [HexNAc]4[Hex]7[Neu5Ac]2
20 + A 1324.3492 RPTGEVYDIEIDTLETTCHVLDPDPLANSVR + [HexNAc]5[Hex]5
4 + B 1324.3458 TLKNLTVETQDAVFVTVELTHPNVK + [HexNAc]4[Hex]5
4 + C 1324.3157 DSVNLTWTEPSDGGSKITNYIVEK + [HexNAc]5[Hex]6[Neu5Ac]2
4 + D 1324.3101 LPPYTPGPSTPWVTNVR + [HexNAc]6[Hex]7[Neu5Ac]3[Fuc]1
21 + A 1415.6259 RPTGEVYDIEIDTLETTCHVLDPDPLANSVR + [HexNAc]5[Hex]6
4 + B 1415.5982 NAAGNFESPSGDSSGATARDEIDAPNALSEDPK + [HexNAc]5[Hex]7[Fuc]2
4 + C 1415.6288 TLKNLTVETQDAVFVTVELTHPNVK + [HexNAc]5[Hex]6
22 + A 1547.0565 VVHAVEVALATFNASENGSYYLQLVEISR + [HexNAc]5[Hex]6
[HexNAc][4][Hex][5]
3 + B 1547.0224 TDTMRLLRPEPLFLPLYNK + [HexNAc][4][Hex][5][Neu5Ac][2]
3 + C 1547.0223 LTPESTREFCNSIHQPQLK^4 + [HexNAc][4][Hex][8]
3 + D 1547.0314 NVTVIEGSVEETCHISGYPSTTTWY^4 + [HexNAc][5][Hex][3]
23 3 + A 1668.7672 VVHAVEVALATFAESNGSYQLVLVEISR + [HexNAc][5][Hex][6]
3 + B 1668.7332 TDTMRLLRPEPLFLPLYNK + [HexNAc][5][Hex][6][Neu5Ac][2]
3 + C 1668.7370 LTPESTREFCNSIHQPQLK + [HexNAc][5][Hex][9]
3 + D 1668.7384 AMKDGVHIDPAQLETANSSSVIIPECK^7 + [HexNAc][4][Hex][4][Neu5Ac][1][Fuc][1]
3 + E 1668.7950 ATAVVEVNLDPGPAPFADITDVTNESCGLTWNP^8 + [HexNAc][2][Hex][4]
24 2 + A 1682.2271 LPCDCPLPLAPLNSDR + [HexNAc][4][Hex][5]
2 + B 1682.1743 SNCTVSVHVSDR^4 + [HexNAc][4][Hex][5][Neu5Ac][1][Fuc][1]
2 + C 1682.2459 LRINRTHASDEGYK + [HexNAc][5][Hex][4]
2 + D 1682.1895 VTNVTK + [HexNAc][7][Hex][7][Fuc][1]
25 2 + A 1864.7932 LPCDCPLPLAPLNSDR + [HexNAc][5][Hex][6]
2 + B 1864.7574 DSGYSLSLTAENSSGTTDQK + [HexNAc][6][Hex][3]
2 + C 1864.7289 AYATITNCTK^4 + [HexNAc][4][Hex][7][Neu5Ac][2]
2 + D 1864.7693 YDSKGYTLTLENSSGTK + [HexNAc][2][Hex][9]
26 4 + A 848.3585 CGLVPVLAENYNK + [HexNAc][4][Hex][5][Neu5Ac][1]
4 + B 848.3761 VNKTIHHDQFK + [HexNAc][4][Hex][7]
4 + C 848.3853 IRDHALDDQANYVSLTNHR + [HexNAc][2][Hex][3][Fuc][1]
27 4 + A 921.1397 CGLVPVLAENYNK + [HexNAc][4][Hex][5][Neu5Ac][2]
4 + B 921.1512 MSDAGKYTVVAGAVNSTAK + [HexNAc][5][Hex][5]
4 + C 921.1133 DGFNITTSEK + [HexNAc][5][Hex][6][Neu5Ac][2]
4 + D 921.1763 LLTQNSNENSEHYYTLVMD + [HexNAc][2][Hex][4]
28 5 + A 944.7902 QQQHLFGSNVTDGCSGNCIFLR + [HexNAc][4][Hex][5][Neu5Ac][2]
5 + B 944.8100 GQVDLVDTMFLAVPNSR + [HexNAc][6][Hex][7][Neu5Ac][1]
5 + C 944.7862 NNTLVLOVR + [HexNAc][6][Hex][7][Neu5Ac][4][Fuc][1]
5 + D 944.8017 NVTFTSUSHMRPPP + [HexNAc][5][Hex][9][Neu5Ac][2]
29 4 + A 1107.9621 QQQHLFGSNVTDGCSGNCIFLR + [HexNAc][4][Hex][5][Neu5Ac][1]
4 + B 1107.9661 VNRNLTVL + [HexNAc][6][Hex][7][Neu5Ac][3][Fuc][1]
4 + C 1107.9878 ANDTTLTSVEPCAGLVELEYVSFR + [HexNAc][3][Hex][6]
4 + D 1107.9765 NVTFTSVIRGTPPP + [HexNAc][5][Hex][9][Neu5Ac][1]
4 + E 1107.9609 NSILWTKVINK + [HexNAc][6][Hex][7][Neu5Ac][3]
30 4 + A 1180.7359 QQQHLFGSNVTDGCSGNCIFLR + [HexNAc][4][Hex][5][Neu5Ac][2]
4 + B 1180.7302 RANHTPECPETKYK + [HexNAc][6][Hex][5][Neu5Ac][3]
4 + C 1180.7608 HILVINDSQFDVGITYAEVEGK + [HexNAc][6][Hex][3][Neu5Ac][1][Fuc][1]
4 + D 1180.7617 ANDTTLTSVEPCAGLVELEYVSFR + [HexNAc][3][Hex][6][Neu5Ac][1]
31 3 + A 1227.8504 CGLVPVLAENYNK + [HexNAc][4][Hex][5][Neu5Ac][2]
3 + B 1227.8153 DGFNITTSEK + [HexNAc][5][Hex][6][Neu5Ac][2]
3 + C 1227.8591 QNATVQQLIQGK + [HexNAc][5][Hex][6][Neu5Ac][1][Fuc][1]
3 +  D  1227.8656  MSDAGKYVVAGGNVSTAK + [HexNAc]5[Hex]5  
3 +  E  1227.9028  ATMRFNTEITAENLTINLK + [HexNAc]5[Hex]3  
32 +  A  1379.9152  QQQHLFGSNVTDCGNFCLFR + [HexNAc]4[Hex]5 
3 +  B  1379.9344  NVTFTSVIRTPPPFK + [HexNAc]5[Hex]9 
3 +  C  1379.9085  NNTLVLQVR + [HexNAc]6[Hex]7[Neu5Ac]2[Neu]1 
3 +  D  1379.8840  VHTNATIR + [HexNAc]6[Hex]7[Neu5Ac]3 
33 +  A  1434.3702  CGLVPVLAENYKSDNEDTPEAGYFAVAVVK + [HexNAc]4[Hex]5[Neu5Ac]2 
4 +  B  1434.3555  LNGSAPIQCVWYRDGVLLR + [HexNAc]6[Hex]7[Neu5Ac]4 
4 +  C  1434.4053  QQHLSFVTPVLAENYNKSDNCEDTPEAGYFAVAVVK + [HexNAc]4[Hex]5[Neu5Ac]3 
4 +  D  1434.3639  DNKEIRPGGNYTITCVGNTPHLR + [HexNAc]7[Hex]6[Neu5Ac]1[Neu]1 
34 +  A  1476.9470  QQQHLFGSNVTDCGSNFCGLFR + [HexNAc]4[Hex]5[Neu5Ac]1 
3 +  B  1476.9160  VHTNATIR + [HexNAc]6[Hex]7[Neu5Ac]4 
3 +  C  1476.9793  HILVINDSQFDDEGVYTAEVEGK + [HexNAc]6[Hex]3[Neu5Ac]1 
3 +  D  1476.9800  GQVDLVDTMAFLVIPNSTR + [HexNAc]6[Hex]7[Neu5Ac]1 
36 +  A  1513.1582  CGLVPVLAENYNK + [HexNAc]4[Hex]5[Neu5Ac]1 
2 +  B  1513.1788  VNKTIIHDTQFK + [HexNAc]3[Hex]6 
2 +  C  1513.1711  QQHLSFVTPVLAENYNK + [HexNAc]4[Hex]5[Neu5Ac]1[Neu]1 
2 +  D  1513.1284  NNVTLK + [HexNAc]6[Hex]6[Neu5Ac]1 
37 +  A  1525.6329  QQQHLFGSNVTDCGSNFCGLFR + [HexNAc]4[Hex]5[Neu5Ac]1[Neu]1 
3 +  B  1525.6292  INNLTESDQGYEVCTEGGTSK + [HexNAc]5[Hex]6 
3 +  C  1525.6659  GQVDLVDTMAFLVIPNSTR + [HexNAc]6[Hex]7[Neu5Ac]1 
3 +  D  1525.6018  VHTNATIR + [HexNAc]6[Hex]7[Neu5Ac]4[Neu]1 
38 +  A  1574.0003  QQQHLFGSNVTDCGNFCLFR + [HexNAc]5[Hex]6[Neu5Ac]1[Neu]1 
3 +  B  1574.0118  GQVDLVDTMAFLVIPNSTR + [HexNAc]6[Hex]7[Neu5Ac]1 
3 +  C  1573.9771  NSILWTKVNK + [HexNAc]6[Hex]7[Neu5Ac]4 
3 +  D  1574.0243  LLQNSENITIENITYHVLVMK + [HexNAc]4[Hex]7[Neu5Ac]1 
39 +  A  1647.3437  QQQHLFGSNVTDCGNFCLFR + [HexNAc]5[Hex]6[Neu5Ac]1[Neu]1 
3 +  B  1647.3598  CGPGPEAYVDEPNMSTAPTVPDPPENVK + [HexNAc]3[Hex]7[Neu5Ac]1 
3 +  C  1647.3421  NSILWTKVNK + [HexNAc]7[Hex]8[Neu5Ac]3[Neu]1 
3 +  D  1647.4079  AWTPYTVTRQVQGLIQGK + [HexNAc]5[Hex]1 
40 +  A  1695.6895  QQQHLFGSNVTDCGSNFCGLFR + [HexNAc]5[Hex]6[Neu5Ac]2 
3 +  B  1695.6829  NNTLVLQVR + [HexNAc]4[Hex]8[Neu5Ac]1 
3 +  C  1695.7014  MSDAGKYVVAGGNVSTAK + [HexNAc]8[Hex]9[Neu5Ac]1 
3 +  D  1695.7225  WVCNFTHDSECQYVTGLSPGDR + [HexNAc]4[Hex]6[Neu5Ac]1[Neu]1 
41 +  A  1792.7213  QQQHLFGSNVTDCGSNFCGLFR + [HexNAc]5[Hex]6[Neu5Ac]2
For each test, candidate A is the actual glycopeptide composition that corresponds to the MS/MS data being scored.

All candidate compositions have an m/z value, calculated in silico, within 50 ppm error of the monoisotopic mass present in the experimental MS data.

The normalization threshold used for determining the presence or absence of the [peptide + core component] product ions in all scoring by GPG was 2 % relative abundance. For Y1 ion evaluation included in the GPG scores reported herein, a relative abundance threshold of 20 % was used for a spectral match to be considered valid.

Normalization thresholds applied to the detection of [precursor – monosaccharide] product ions vary according to identity of the neutral loss being evaluated.

Denotes peptides where Cys residues were not derivatized by IAM.

3.3.3 GPG Validation: Application to Recombinant Gp120 HIV Envelope

Glycoprotein. As the GPG algorithm was designed after studying the fragmentation patterns obtained for RNase B, asialofetuin, and transferrin, whose spectra comprise the training data set, it was expected that the automated GPG tool would perform well when testing the training data set. Therefore, after analysis of the training data set, the GlycoPep Grader software was used to
analyze CID data collected on tryptic digests of the HIV envelope protein, CON-S gp140 CFI. The resulting CID spectra from the CON-S gp 140 CFI glycopeptides (herein referred to as the validation data set) contain MS/MS data on glycopeptides of varying N-linked glycan types and compositional arrangements. A total of over 100 CID spectra from 34 unique CON-S gp140 CFI glycopeptides were tested using the GPG tool. These results are summarized in Table 3.
Table 3. GPG Score Results of Tests Performed on CON-S gp140 CFI Glycopeptide CID Spectra Comprising the Validation Data Set.

<table>
<thead>
<tr>
<th>Test</th>
<th>Charge</th>
<th>Candidate</th>
<th>m/z</th>
<th>Glycopeptide Composition</th>
<th>GPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 +</td>
<td>A</td>
<td>1071.5123</td>
<td>SNITGLLITR + [HexNAc]2[Hex]4</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>B</td>
<td>1071.4758</td>
<td>ANVTVEAR + [HexNAc]4[Hex]2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>C</td>
<td>1071.4695</td>
<td>MANISRYYEEP + [HexNAc]2[Hex]2</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>3 +</td>
<td>A</td>
<td>1156.4722</td>
<td>DGGNNNTNETEIFRPGGDMR + [HexNAc]2[Hex]5</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>B</td>
<td>1156.5238</td>
<td>YTLVENNSGSKSITFTVK + [HexNAc]2[Hex]6</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>C</td>
<td>1156.4895</td>
<td>VDGHEWTCKNTTPTK + [HexNAc]4[Hex]5</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>D</td>
<td>1156.5225</td>
<td>NNLPISSNVSISR + [HexNAc]6[Hex]4</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2 +</td>
<td>A</td>
<td>1160.4919</td>
<td>EANNTLFCASDAK + [HexNAc]2[Hex]3</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>B</td>
<td>1160.4997</td>
<td>ANVTVEAR + [HexNAc]4[Hex]4</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>C</td>
<td>1160.5104</td>
<td>YQSNATLVCK + [HexNAc]4[Hex]2</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>D</td>
<td>1160.4567</td>
<td>ANK + [HexNAc]5[Hex]6</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>2 +</td>
<td>A</td>
<td>1193.5653</td>
<td>SNITGLLITR + [HexNAc]4[Hex]3</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>B</td>
<td>1193.5372</td>
<td>LNWTKEHDDGAGK + [HexNAc]3[Hex]2</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>C</td>
<td>1193.5239</td>
<td>YQSNATLPQGN + [HexNAc]4[Hex]2</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>2 +</td>
<td>A</td>
<td>1254.0784</td>
<td>SNITGLLITR + [HexNAc]3[Hex]5</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>B</td>
<td>1254.0502</td>
<td>LNWTKEHDDGAGK + [HexNAc]2[Hex]4</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>C</td>
<td>1254.0503</td>
<td>ANVTVEAR + [HexNAc]5[Hex]3[Fuc]1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>D</td>
<td>1254.0399</td>
<td>YCVVVENSTGSR + [HexNAc]4[Hex]2</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>3 +</td>
<td>A</td>
<td>1264.5074</td>
<td>DGGNNNTNETEIFRPGGDMR + [HexNAc]2[Hex]7</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>B</td>
<td>1264.5590</td>
<td>YTLVENNSGSKSITFTVK + [HexNAc]2[Hex]8</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>C</td>
<td>1264.5247</td>
<td>VDGHEWTCKNTTPTK + [HexNAc]4[Hex]7</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>D</td>
<td>1264.5171</td>
<td>NASGSAAKEIK + [HexNAc]5[Hex]6[Neu5Ac]2[Fuc]1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>3 +</td>
<td>A</td>
<td>1268.8550</td>
<td>NNNNTNTDITLTPCR + [HexNAc]6[Hex]4[Neu5Ac]1</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>B</td>
<td>1268.8474</td>
<td>YTCQAKNESGVER + [HexNAc]5[Hex]5[Neu5Ac]1[Fuc]1</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>C</td>
<td>1268.8582</td>
<td>QSDAGEYTFVAGRN + [HexNAc]5[Hex]6[Fuc]1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>D</td>
<td>1268.8608</td>
<td>TKANVTVEAR + [HexNAc]5[Hex]6[Neu5Ac]2[Fuc]1</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>E</td>
<td>1268.8474</td>
<td>YTCQAKNESGVER + [HexNAc]5[Hex]5[Neu5Ac]1[Fuc]1</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>F</td>
<td>1268.8710</td>
<td>ANKTPIRRM + [HexNAc]7[Hex]8</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>3 +</td>
<td>A</td>
<td>1272.8513</td>
<td>DGGNNNTNETEIFRPGGDMR + [HexNAc]3[Hex]5[Fuc]1</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>B</td>
<td>1272.8834</td>
<td>VRDNDAGNFTCRATNSVGSK + [HexNAc]5[Hex]3[Fuc]1</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>C</td>
<td>1272.8393</td>
<td>VRDNDAGNFTCR + [HexNAc]5[Hex]4[Neu5Ac]2[Fuc]1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>D</td>
<td>1272.9008</td>
<td>KCSTKFVENLTGAIWYFR + [HexNAc]2[Hex]6</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>3 +</td>
<td>A</td>
<td>1286.5269</td>
<td>DGGNNNTNETEIFRPGGDMR + [HexNAc]4[Hex]4[Fuc]1</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>B</td>
<td>1286.5201</td>
<td>QNLTVDVTKV + [HexNAc]7[Hex]8[Neu5Ac]1[Fuc]1</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>C</td>
<td>1286.5467</td>
<td>INGSEPLQSVWYK + [HexNAc]6[Hex]6[Fuc]1</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>D</td>
<td>1286.5430</td>
<td>MSDAGKYTVVAGGNVSTAK + [HexNAc]3[Hex]5[Neu5Ac]2</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>2 +</td>
<td>A</td>
<td>1298.5581</td>
<td>NCSNFITTEIR + [HexNAc]3[Hex]3[Fuc]1</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>B</td>
<td>1298.5734</td>
<td>YTLTLSENSGTK + [HexNAc]4[Hex]2[Fuc]1</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>C</td>
<td>1298.5453</td>
<td>YTCQAKNESGVER + [HexNAc]2[Hex]4</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>2 +</td>
<td>D</td>
<td>1298.5733</td>
<td>YLTVENNSGSK + [HexNAc]4[Hex]3</td>
<td>29</td>
</tr>
<tr>
<td>11</td>
<td>3 +</td>
<td>A</td>
<td>1318.5250</td>
<td>DGGNNNTNETEIFRPGGDMR + [HexNAc]2[Hex]8</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>3 +</td>
<td>B</td>
<td>1318.5804</td>
<td>CSKTSFKVENLTAGIYYFR + [HexNAc]2[Hex]7</td>
<td>80</td>
</tr>
<tr>
<td>Charge</td>
<td>Protein Sequence</td>
<td>Mass (Da)</td>
<td>HexNAc</td>
<td>Hex</td>
<td>Neu5Ac</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------</td>
<td>-----------</td>
<td>--------</td>
<td>-----</td>
<td>--------</td>
</tr>
<tr>
<td>3</td>
<td>EKNSILWVKLNK + [HexNAc]6[Hex]6[Neu5Ac]1</td>
<td>1318.5888</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>VQIEKGVNQYQLSINDNCNDRDAGK + [HexNAc]2[Hex]5</td>
<td>1318.5902</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>KAYANVSSKCSK + [HexNAc]6[Hex]5[Neu5Ac]1</td>
<td>1318.5457</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>NCSFNTITEIR + [HexNAc]4[Hex]3</td>
<td>1327.0688</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>FTNITGEK + [HexNAc]3[Hex]7</td>
<td>1327.0412</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>LNWTKEHDGGAK + [HexNAc]2[Hex]4[Hex]1</td>
<td>1327.0793</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>RGRQNLTVK + [HexNAc]3[Hex]6</td>
<td>1327.0527</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>VQIEKGVNYTQLSIDNCDRNDAGK + [HexNAc]2[Hex]5</td>
<td>1327.0767</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>NAYANVSSKCSK + [HexNAc]6[Hex]5[Neu5Ac]2</td>
<td>1327.0528</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>NCSFNITTEIR + [HexNAc]3[Hex]5</td>
<td>1335.0606</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>NCSFNITTEIR + [HexNAc]2[Hex]6</td>
<td>1335.0765</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>VQIEKGVNYTQLSIDNCDRNDAGK + [HexNAc]2[Hex]5</td>
<td>1335.0257</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>NCSFNITTEIR + [HexNAc]3[Hex]5</td>
<td>1335.0764</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>ANVTVEAR + [HexNAc]5[Hex]5</td>
<td>1335.0284</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]9</td>
<td>1343.0586</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>CSKTSFKVENLTEGAIYYFR + [HexNAc]4[Hex]2</td>
<td>1343.0658</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>ANVTVEAR + [HexNAc]5[Hex]5</td>
<td>1343.0781</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>ANVTVEAR + [HexNAc]5[Hex]5</td>
<td>1343.0764</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>NCSFNITTEIR + [HexNAc]3[Hex]5</td>
<td>1351.0686</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTFYAGENITSGK + [HexNAc]4[Hex]2</td>
<td>1351.0545</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTLTLENSSTGK + [HexNAc]3[Hex]5</td>
<td>1351.0839</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>DGRQNLTVK + [HexNAc]2[Hex]6</td>
<td>1351.0764</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]9</td>
<td>1367.0686</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>CSKTSFKVENLTEGAIYYFR + [HexNAc]4[Hex]2</td>
<td>1367.0545</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTFYAGENITSGK + [HexNAc]4[Hex]2</td>
<td>1367.0867</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTLTLENSSTGK + [HexNAc]3[Hex]5</td>
<td>1367.0839</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>DGRQNLTVK + [HexNAc]2[Hex]6</td>
<td>1367.0764</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]9</td>
<td>1387.5819</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>CSKTSFKVENLTEGAIYYFR + [HexNAc]4[Hex]2</td>
<td>1387.5980</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTFYAGENITSGK + [HexNAc]4[Hex]2</td>
<td>1387.6064</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTLTLENSSTGK + [HexNAc]3[Hex]5</td>
<td>1387.6039</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>DGRQNLTVK + [HexNAc]2[Hex]6</td>
<td>1387.5896</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]9</td>
<td>1400.0978</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>CSKTSFKVENLTEGAIYYFR + [HexNAc]4[Hex]2</td>
<td>1400.0996</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTFYAGENITSGK + [HexNAc]4[Hex]2</td>
<td>1400.1051</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTLTLENSSTGK + [HexNAc]3[Hex]5</td>
<td>1400.1016</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>DGRQNLTVK + [HexNAc]2[Hex]6</td>
<td>1400.1312</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]9</td>
<td>1416.0978</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>CSKTSFKVENLTEGAIYYFR + [HexNAc]4[Hex]2</td>
<td>1416.1096</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTFYAGENITSGK + [HexNAc]4[Hex]2</td>
<td>1416.1051</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]9</td>
<td>1451.2518</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>CSKTSFKVENLTEGAIYYFR + [HexNAc]4[Hex]2</td>
<td>1451.2582</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTFYAGENITSGK + [HexNAc]4[Hex]2</td>
<td>1451.2582</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>YTLTLENSSTGK + [HexNAc]3[Hex]5</td>
<td>1451.2679</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

98
| 3 + E | 1451.2451 | LVINR + [HexNAc][Hex][Neu5Ac][Fuc] | 7 |
| 3 + F | 1451.2328 | NVTLK + [HexNAc][Hex][Neu5Ac][Fuc] | 6 |
| 22 2 + A | 1436.6003 | EANTTLFCASDAK + [HexNAc][Hex][Fuc] | 81 |
| 22 2 + B | 1436.5998 | YCVVVENSTGSR + [HexNAc][Hex][Neu5Ac] | 40 |
| 22 2 + C | 1436.6186 | RANHTPESCPETKYK + [HexNAc][Hex][Neu5Ac] | 8 |
| 22 2 + D | 1436.6163 | LNWTPEHDDGAK + [HexNAc][Hex][Neu5Ac] | 17 |
| 23 2 + A | 1476.6443 | SNITGLLTLTR + [HexNAc][Hex][Neu5Ac] | 82 |
| 23 2 + B | 1476.6240 | YTFYAGENITSGK + [HexNAc][Hex][Neu5Ac] | 0 |
| 23 2 + C | 1476.6079 | ANVTVEAR + [HexNAc][Hex][Neu5Ac] | 0 |
| 23 2 + D | 1476.6321 | NGINTPSQR + [HexNAc][Hex][Neu5Ac] | 3 |
| 23 2 + E | 1476.6123 | YQSNATLVCK + [HexNAc][Hex][Neu5Ac] | 17 |
| 24 2 + A | 1481.1242 | NCSFNTIITEIR + [HexNAc][Hex][Neu5Ac][Fuc] | 80 |
| 24 2 + B | 1481.1576 | YILLKLENSGSK + [HexNAc][Hex][Neu5Ac] | 52 |
| 24 2 + C | 1481.1687 | QNATQVQLGQKG + [HexNAc][Hex][Neu5Ac] | 39 |
| 24 2 + D | 1481.1392 | YTVVAGGNVSTAK + [HexNAc][Hex][Neu5Ac][Fuc] | 11 |
| 24 2 + E | 1481.1443 | NGTEILKSK + [HexNAc][Hex][Neu5Ac][Fuc] | 32 |
| 24 2 + F | 1481.1423 | VENLATEGAYYFR + [HexNAc][Hex][Neu5Ac] | 0 |
| 24 2 + G | 1481.1395 | YITLSENSGTG + [HexNAc][Hex][Neu5Ac][Fuc] | 34 |
| 24 2 + H | 1481.1394 | YITLVENSSGSK + [HexNAc][Hex][Neu5Ac] | 44 |
| 24 2 + I | 1481.1315 | TKANVTVEAR + [HexNAc][Hex][Neu5Ac][Fuc] | 0 |
| 24 2 + J | 1481.1889 | INETLELSESPVYSTK + [HexNAc][Hex][Neu5Ac][Fuc] | 68 |
| 25 2 + A | 1517.6262 | EANNTTLFCASDAK + [HexNAc][Hex][Neu5Ac][Fuc] | 72 |
| 25 2 + B | 1517.6144 | ANHTPESCPETK + [HexNAc][Hex][Neu5Ac][Fuc] | 29 |
| 25 2 + C | 1517.6162 | ANDTLVR + [HexNAc][Hex][Neu5Ac][Fuc] | 6 |
| 25 2 + D | 1517.6388 | YQSNATLVCK + [HexNAc][Hex][Neu5Ac][Fuc] | 40 |
| 26 2 + A | 1568.1958 | LINCNTSAITQACPK + [HexNAc][Hex][Neu5Ac][Fuc] | 100 |
| 26 2 + B | 1568.1583 | NASGSKAKEI + [HexNAc][Hex][Neu5Ac][Fuc] | 11 |
| 26 2 + C | 1568.2197 | RESGTTAWQVNSSVKR + [HexNAc][Hex][Neu5Ac][Fuc] | 15 |
| 26 2 + D | 1568.1898 | LENSIGSKAFVTVK + [HexNAc][Hex][Neu5Ac][Fuc] | 17 |
| 27 2 + A | 1583.1647 | LDVPIDDDNSSNYR + [HexNAc][Hex][Neu5Ac][Fuc] | 100 |
| 27 2 + B | 1583.1998 | VRNLNVTLK + [HexNAc][Hex][Neu5Ac][Fuc] | 32 |
| 27 2 + C | 1583.1425 | VETCNLSEVEK + [HexNAc][Hex][Neu5Ac][Fuc] | 0 |
| 27 2 + D | 1583.1930 | AYATITNCTKTFR + [HexNAc][Hex][Neu5Ac][Fuc] | 0 |
| 28 2 + A | 1603.1722 | NCSFNTIITEIR + [HexNAc][Hex][Neu5Ac][Fuc] | 78 |
| 28 2 + B | 1603.1885 | TCEILEILNSTK + [HexNAc][Hex][Neu5Ac][Fuc] | 7 |
| 28 2 + C | 1603.1644 | YTCQAKNESGER + [HexNAc][Hex][Neu5Ac][Fuc] | 46 |
| 28 2 + D | 1603.1973 | QNLTLDVTKV + [HexNAc][Hex][Neu5Ac][Fuc] | 21 |
| 29 2 + A | 1610.1478 | NCSFNTIITEIR + [HexNAc][Hex][Neu5Ac][Fuc] | 89 |
| 29 2 + B | 1610.1877 | VFAENETGLSRPR + [HexNAc][Hex][Neu5Ac][Fuc] | 25 |
| 29 2 + C | 1610.1661 | YTFYAGENITSGK + [HexNAc][Hex][Neu5Ac][Fuc] | 11 |
| 29 2 + D | 1610.1137 | NASGTK + [HexNAc][Hex][Neu5Ac][Fuc] | 13 |
| 30 2 + A | 1669.7355 | LINCNTSAITQACPK + [HexNAc][Hex][Neu5Ac][Fuc] | 95 |
| 30 2 + B | 1669.6986 | NDAGKYTLVENSNSGK + [HexNAc][Hex][Neu5Ac][Fuc] | 5 |
| 30 2 + C | 1669.7664 | DTGEYTLLEKNVTGTTSETIK + [HexNAc][Hex][Neu5Ac][Fuc] | 42 |
| 30 2 + D | 1669.6873 | WVRHNK + [HexNAc][Hex][Neu5Ac][Fuc] | 24 |
| 31 2 + A | 1697.2203 | LDVPIDDDNSSNYR + [HexNAc][Hex][Neu5Ac][Fuc] | 95 |
| 31 2 + B | 1697.2139 | TNKTIHNDTQFK + [HexNAc][Hex][Neu5Ac][Fuc] | 52 |
For each test, candidate A is the actual glycopeptide composition that corresponds to the MS/MS data being scored. All candidate compositions have an m/z value, calculated in silico, within 50 ppm error of the monoisotopic mass present in the experimental MS data. The normalization threshold used for determining the presence or absence of the [peptide + core component] product ions in all scoring by GPG was 2% relative abundance. For Y ion evaluation included in the GPG scores reported herein, a relative abundance threshold of 20% was used for a spectral match to be considered valid. Normalization values applied to the detection of [precursor – monosaccharide] product ions vary according to the identity of the neutral loss being evaluated. Denotes peptides where Cys residues were not derivatized by IAM.

A minimum of three candidate compositions were scored for each spectrum, with an average of four to five glycopeptide candidates being evaluated in each test performed. In agreement with the training data set results, the GPG algorithm assigned the highest score to each correct candidate composition, for each CON-S gp140ΔCFI glycopeptide spectrum, scored in the validation data set. An example of a scored fucosylated complex type structure from CON-S gp140 ΔCFI is shown in Figure 3.
Figure 3. MS\(^2\) data from the validation set. (A) GPG evaluation of the correct candidate
Figure 3. MS² data from the validation set. (A) GPG evaluation of the correct candidate composition assignment for a fucosylated N-glycopeptide from CON-S gp140 CFI. Scoring in (B) and (C) shows evaluation of this spectrum against decoy candidate compositions with the same nominal mass. Arrows with an X indicate ions that were not present in the spectra. A 2 % relative abundance threshold was used for [peptide + core component] product ion matching to decrease false positives from noise. For the composition in (A), GPG generated a score of 95 %. The decoy compositions in (B) and (C) were scored at 52 % and 68 % respectively, indicating that GPG scored the correct compositional assignment as the most probable glycopeptide composition from this pool of candidates.

The GPG scores for decoy compositions tested against this spectrum are also reported on the spectra. This example is Test 31 of Table 3.

While the data from both the training sets and validation sets were quite encouraging, one might note that in each case, a limited number of decoys were tested against the true composition. To test the likelihood that this limited number of decoys was a required feature for the correct candidate to get the top score, a glycopeptide spectrum from gp140 was tested against nine alternate isobaric candidate compositions. Scores are shown in Table 1, and the MS/MS data is in Figure 4.
Figure 4. CID spectrum of a glycopeptide from CON-S gp140 CFI scored by GPG, along with nine alternate compositions, as shown in Table 1. Even when an extensive number of candidates share the same nominal mass, GPG scored the correct composition as the most probable glycopeptide match for the MS$^2$ data against all other potential assignments tested. The correct composition of NCSFNITTEIR + [HexNAc]$_4$[Hex]$_4$[Fuc]$_1$ was indicated with the highest GPG score, 80 %, while the highest scoring decoy composition, INETLELLSEPVYSTK + [HexNAc]$_2$[Hex]$_3$[Fuc]$_1$ was assigned a GPG score of 68 %.
Table 4. GPG Results for Candidate Compositions Tested Against the MS\(^2\) Data in Figure 4.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Mass (Da)</th>
<th>Glycopeptide Composition</th>
<th>Score (^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2960.2338</td>
<td>NCSFNITTEIR + [HexNAc]4[Hex]4[Fuc]1</td>
<td>80</td>
</tr>
<tr>
<td>B</td>
<td>2960.3006</td>
<td>YILKLENSGSK + [HexNAc]4 + [Hex]5</td>
<td>52</td>
</tr>
<tr>
<td>C</td>
<td>2960.3228</td>
<td>QNATVQGLIQGK + [HexNAc]6[Hex]3</td>
<td>39</td>
</tr>
<tr>
<td>D</td>
<td>2960.2638</td>
<td>YTVVAGGNVSTAK + [HexNAc]3[Hex]4[Neu5Ac]1[Fuc]1</td>
<td>11</td>
</tr>
<tr>
<td>E</td>
<td>2960.2740</td>
<td>NGTEILKSK + [HexNAc]5[Hex]5[Fuc]1</td>
<td>32</td>
</tr>
<tr>
<td>F</td>
<td>2960.2700</td>
<td>VENLTEGAIYYFR + [HexNAc]3[Hex]3[Neu5Ac]1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>2960.2644</td>
<td>YTLTLLENSSGKT + [HexNAc]5[Hex]3[Fuc]1</td>
<td>34</td>
</tr>
<tr>
<td>H</td>
<td>2960.2642</td>
<td>YILTVENSSGSK + [HexNAc]5[Hex]4</td>
<td>44</td>
</tr>
<tr>
<td>I</td>
<td>2960.2484</td>
<td>TKANVTVEAR + [HexNAc]3[Hex]6[Neu5Ac]1</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>2960.3632</td>
<td>INETLELLSESPVYSTK + [HexNAc]2[Hex]3[Fuc]2</td>
<td>68</td>
</tr>
</tbody>
</table>

\(^1\) Candidate A is the actual \(N\)-linked glycopeptide composition corresponding to the CID spectrum scored by GPG and candidates B, C, D, E, F, G, H, I, and J are decoy compositions of nearly identical neutral mass.

\(^2\) All glycopeptide compositions have an \(m/z\) value, calculated \(in\) \(silo\), that is within 50 ppm error of the monoisotopic mass present in the experimental MS\(^1\) data. Users may also utilize low resolution MS\(^1\) data to determine glycopeptide candidates, though more compositions will result.

\(^3\) Denotes GPG scores at 2 % peptide normalization.

Although the score values and distribution varies from spectrum to spectrum, GPG ranked the correct candidate composition as the most probable glycopeptide in each test performed, including approximately 300 CID spectra from the training and validation sets. A screen shot of the GPG scoring output for a high mannose type CON-S gp140 CFI glycopeptide, along with three decoy candidate compositions, is included in Figure 5.
Figure 5. Screen shot of output showing GPG scoring of a glycopeptide from CON-S gp140 CFI, along with three candidate compositions with a nearly identical neutral mass. The actual glycopeptide composition of SNITGLLLETR + [HexNAc]2[Hex]8 received a score of 79 %, while the three decoy compositions of NGINVTPSQR + [HexNAc]6[Hex]3, DNGSPILGYWLEK + [HexNAc]4[Hex]3, and ANDTLVR + [HexNAc]3[Hex]5[Neu5Ac]2 received scores of 3 %, 4 %, and 6 %, respectively. The exact m/z values, calculated in silico, are shown for each of the glycopeptide candidate compositions in Table 3, test #18.

3.4 CONCLUDING REMARKS

We have developed a novel software analysis tool, GlycoPep Grader, to increase the speed and efficiency of assigning N-linked glycopeptide composition from MS/MS data. This novel spectral scoring approach relies heavily on the identification of the peptide-containing, or [peptide + core component], product ions and neutral monosaccharide residue losses, or [precursor – monosaccharide] product ions, across various charge states. After developing and testing the GPG software using a training set of CID data collected on glycopeptides from RNase B, asialofetuin, and transferrin, GPG was then validated by scoring glycopeptide compositions from the recombinant HIV envelope protein, CON-S gp140 CFI, against alternate candidate compositions of the same nominal mass. Thus far, in approximately 300 tests performed across spectra of differing quality, the novel scoring algorithm powering GPG identifies the correct glycopeptide composition as the highest scoring candidate ion every time.

This tool has several useful features, compared to other existing glycopeptide analysis tools: 1) It is the only available tool whose scoring algorithm was designed specifically for low resolution CID data, 2) It does not require the user to first deconvolute the spectrum to singly
charged ions, which is often difficult or impossible for low resolution CID spectra, 3) It has unique scoring rules, depending on the types of glycans present in the candidate composition and 4) The user need not know the peptide composition in advance in order to use the tool, but rather inputs potential candidate compositions obtained from available glycopeptide databases that correspond to the precursor’s experimental mass. Finally, GPG has shown unprecedented success in accurately identifying of the correct glycopeptide composition in 79 unique test cases.

3.5 ACKNOWLEDGEMENTS

The author acknowledges financial support from the NIH (RO1RR026061) to H.D., an NSF Fellowship (DGE-0742523) to C.W. and K.R., and a Pfizer Scholarship to C.W.

The author also wishes to thank all co-authors on the GPG publication for their effort and contribution in making the success of this project possible: David Hua for his time and work, especially in writing the GPG software code, Morgan Maxon for her work in the development and testing of GPG candidate compositions, Katie Rebecchi for collecting and providing some of the CID spectra collected on the model glycopeptides, Eden Go for supplying the glycopeptide CID spectra collected on CON-S gp140 CFI, and Heather Desaire for her continued mentoring and support.
3.6 REFERENCES


(28) Gao, F.; Weaver, E. A.; Lu, Z.; Li, Y.; Liao, H. X.; Ma, B.; Alam, S. M.; Scearce, R. M.;


CHAPTER 4

COMPUTATIONAL METHOD TO DETERMINE PRECURSOR CHARGE STATE IN ETD MS/MS DATA OF DISULFIDE-BONDED PEPTIDES

ABSTRACT

The analysis of peptides using electron transfer dissociation (ETD) is still in the early stages, as is the development of methods and programs to elucidate a precursor’s charge state from peptide ETD spectra. Conversely, manual assignment of spectra is tedious and time-consuming. Still, accurate charge state assignment is necessary in order to determine mass of a precursor ion. As low resolution instruments, such as the Thermo Scientific Velos LTQ, are equipped with ETD capabilities, the availability of computational tools to assist in the determination of precursor charge state directly from ETD MS/MS data is essential to advance the study of disulfide-bonded peptides.

Although a few programs determine charge state from low resolution ETD data of peptide precursors, the majority are not freely available to the public. Additionally, no program has been described or reported to be tested on peptides containing disulfide bonds. To address this need in automated MS/MS analysis, we have developed a method that utilizes simple computational tools generated in Excel in order to identify charge state in disulfide-bonded peptide precursors. One benefit of the computational tools is that the most likely precursor charge state may be deciphered when more than one potential charge state exists, which greatly reduces the amount of time it takes to perform subsequent protein database searches.
4.1 INTRODUCTION

Disulfide bonds play a critical role in stabilizing protein structure. Experimental studies show that in the absence of proper disulfide bond orientation, proper protein folding may not occur. As such, it is important to identify the disulfide bond arrangements in order to evaluate structural features within a protein. This described importance is magnified within the pharmaceutical and biotechnology industries, where the mapping of disulfide bonds is critical to ensuring drug quality and efficacy. In the characterization of protein disulfide bond formation, analysis by electron transfer dissociation tandem mass spectrometry (ETD MS/MS) is emerging as a powerful technique.

At present, mass spectrometry is a common analysis route used to identify disulfide bond arrangements in proteins and peptides. Nuclear magnetic resonance (NMR) and crystallography experiments can also be used to obtain disulfide bond information; unfortunately, the instrumentation requires high amounts of high purity samples. In contrast, LC-MS reveals disulfide bond patterns using small sample amounts, even when the sample is comprised of unknowns. To obtain experimental MS/MS data on disulfide-bonded peptides, a variety of fragmentation techniques may be utilized, including collision induced dissociation (CID). However, ETD MS/MS impart more extensive fragmentation information for these species.

During ETD MS/MS experiments, Cα-N bond cleavage is induced through the transfer of an electron from a radical anion, such as fluoranthene, to a protonated peptide. The mechanism of peptide backbone cleavage is analogous to the way in which fragmentation in electron capture dissociation (ECD) occurs. When the backbone is fragmented, it dissociates into c- and z-type product ions. Like the b- and y-type ions generated
During collision induced dissociation (CID),\textsuperscript{11, 27, 30} these product ions provide peptide sequence information. However, in the study of disulfide-bonded peptides, ETD presents several advantages over the more popular CID. For example, although ETD generally preserves labile post-translational modifications (PTMs),\textsuperscript{11, 27, 30} it is shown to preferentially cleave between peptides containing a disulfide bond.\textsuperscript{10, 12, 13, 30} This provides valuable MS/MS information about the individual chains that are bonded together to form a disulfide. In contrast, although CID is known to fragment labile PTMs,\textsuperscript{11, 27, 30} it does not typically fragment the covalent disulfide bond.\textsuperscript{10, 11}

Still, in order to utilize low resolution ETD MS instruments, the charge state of the precursor ion must be known. This is a requirement for accurately determining mass, or for identifying peptide sequence using automated database search tools such as MassMatrix.\textsuperscript{32, 33} For example, when the charge state is unknown, repeated searches must be performed on each spectrum to evaluate all possible charge states. These repeated searches result in analysis that is inefficient and time-consuming, which becomes even more costly when larger peptides with a higher charge state distribution are considered.

In the determination of precursor charge state directly from peptide ETD MS/MS data, a few software tools have recently been described in the literature.\textsuperscript{34, 35, 36} Two of these programs are accessible to only select users. These include the commercially available Charger, developed by Sadygov \textit{et al.} and distributed by Thermo Scientific, and Charge Prediction Machine (CPM) by Carvalho \textit{et al.}, which is available to academic users to run on Linux through the Mono Project.\textsuperscript{35, 36} The other program, developed by Sharma \textit{et al.}, utilizes support vector machine (SVM) classifiers to deduce charge state from analyzing patterns in the intensity of the charge reduced precursor ion peaks within an ETD spectrum.\textsuperscript{34} This program is publicly available,
however, it requires the use of other stand-alone programs, and is not intuitive to users unfamiliar with vector analysis.

Although programs such as Charger, CPM, and the prediction tool using SVM classifiers have been created to determine precursor ion charge state directly from low resolution ETD MS/MS data, the fragmentation behavior of peptides containing disulfide bonds has shown different predominant characteristics in comparison to peptides lacking these modifications. For instance, limited fragmentation is observed and not as many c- and z-type product ions are produced, as a majority of the ETD activation energy goes into cleaving the S-S bond. As it stands, no program has been designed to analyze ETD MS/MS data of peptides containing inter and intra disulfide bonds, nor has testing using the above automated programs been reported on them.

We present herein a simple computational method that allows for the determination of precursor ion charge state directly from MS/MS data of disulfide-bonded peptides. This method is applicable to peptides containing both interchain and intrachain type disulfide bonds. In addition to being developed specifically for the interpretation of peptides containing disulfide bonds, the novel computational method utilizes Excel-based tools to deduce the single most probable charge state for a precursor ion.

4.2 EXPERIMENTAL

4.2.1 Materials and Reagents. Chicken lysozyme, bovine fetuin, bovine serum albumin (BSA), human apo-transferrin (transferrin), formic acid, and acetic acid were purchased from Sigma Aldrich (St. Louis, MO). HPLC grade methanol (CH$_3$ OH) and HPLC grade acetonitrile (CH$_3$ CN) were purchased from Fisher Scientific (Fairlawn, NJ). Ammonium bicarbonate (NH$_4$HCO$_3$) was purchased from Fluka (Milwaukee, WI) and sequencing grade modified trypsin.
was from purchased Promega (Madison, WI). Ultrapure water was obtained from a Millipore Direct-Q® UV 3 system (Billerica, MA) with a resistance greater than 18 MΩ.

**4.2.2 Protease Digestion.** Lysozyme, fetuin, transferrin, and BSA were subjected to proteolytic digestion. Approximately 400 µg of each protein was dissolved in 100 mM NH$_4$HCO$_3$ (pH 8.0). Next, trypsin was added to in a 1:80 (w/w) protease to protein ratio and incubated at 37 °C for 24 hr. The protease digestion was stopped by the addition of 1 µL concentrated acetic acid for every 100 µL solution. These samples were then analyzed by LC-MS and subjected to MS/MS experiments, as described below.

**4.2.3 Mass Spectrometry on an ESI-LTQ Velos.** Peptides obtained from each tryptic digest were loaded onto a C18 column (300 µm i.d., 5 cm length, and 3 µm particle size) produced by CVC Microtech (Fontana, CA) at a final concentration of 15 µM after dilution with ultrapure water. The column was connected to a Waters Acquity UPLC system (Milford, MA), directly coupled to an electrospray-linear ion trap mass spectrometer (ESI-LTQ Velos MS) from ThermoScientific (San Jose, CA). The aqueous mobile phase was comprised of 99.9 % water and 0.1 % formic acid (solvent A) and the organic mobile phase consisted of 99.9 % acetonitrile and 0.1 % formic acid (solvent B). The flow rate was set to 7 µL/min. For reversed phase separation of peptides, the solvent conditions were as follows: 2 min at 2 % B, a 10 min linear increase to 5 % B, another 10 min linear increase to 20 % B, a 20 min linear increase to 50 % B, another 20 min linear increase to 60 % B, and a final 10 min linear increase to 95 % B before being held at 95 % B for 10 min, followed by re-equilibration of the column. A 45 min wash cycle and blank injection were used between each sample run to ensure no carry over between samples occurred. For mass spectrometry, the electrospray source voltage was 3 kV and the capillary temperature was 250 °C. For ETD MS/MS analysis of peptides, fluoranthene was used.
as the ETD reagent. An activation time of 100 ms, and an isolation width of 2.5 Da was used.

Supplemental activation was enabled. LC-MS/MS was set up in data dependent scan mode where the 5 most intense ions were chosen for MS/MS analysis with a 3 min dynamic exclusion window. MS/MS scans were collected in centroid mode.

4.2.4 Manual Data Analysis. To identify the peptides containing disulfide bonds in the MS data, a prediction table of theoretical \( m/z \) values corresponding to disulfide-bonded peptides for each of the protein digest samples was prepared. The amino acid sequences from lysozyme, fetuin, transferrin, and BSA were obtained from Uniprot (www.uniprot.org) and imported into Protein Prospector (http://prospector.ucsf.edu/prospector/mshome.htm), where a theoretical tryptic digest was performed. Tryptic miscleavages were not considered. The masses of the peptides that contained both inter-and intra-bonded disulfides were calculated with their modification. These masses were converted into \( m/z \) values corresponding to the disulfide-bonded peptides in multiple charge states. The MS/MS data for lysozyme, fetuin, BSA, and transferrin were then searched to identify spectra that corresponded to the correct \( m/z \) value of the calculated species. The ETD spectra were carefully evaluated in order to verify each assignment made.

4.3 RESULTS AND DISCUSSION

Various proteins with intact disulfide linkages were digested with trypsin and analyzed by ETD-MS/MS in order to develop an automated approach for the determination of precursor charge state directly from their tandem mass spectra. Resultant peptides chosen for charge state determination ranged in length from 2 to 61 amino acids, and contained an assortment of disulfide bond arrangements. Information on the selected proteins is given in Table 1.
### Table 1. Proteins Containing Disulfide Bonds Analyzed by ETD MS/MS.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Mass (Da)</th>
<th>Length (AA)</th>
<th># Cys Residues</th>
<th># Disulfide Bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lysozyme</td>
<td>16,239</td>
<td>147</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Fetuin</td>
<td>38,419</td>
<td>359</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>BSA</td>
<td>69,293</td>
<td>607</td>
<td>35</td>
<td>17</td>
</tr>
<tr>
<td>Transferrin</td>
<td>77,064</td>
<td>698</td>
<td>40</td>
<td>19</td>
</tr>
</tbody>
</table>

1Lysozyme has 9 Cys located within the protein sequence, with 1 Cys in signal peptide and 8 Cys in protein chain. Fetuin has 14 Cys located within the protein sequence, with 2 Cys in signal peptide and 12 Cys in protein chain. BSA has 35 Cys located within the protein sequence, with 0 Cys in signal peptide and 35 Cys in protein chain. Transferrin has 40 Cys located within the protein sequence, with 2 Cys in signal peptide and 38 Cys in protein chain.

4.3.1 Low Resolution ETD MS/MS Data of Peptides Containing Disulfide Bonds. In low resolution ETD MS data, the charge state of a precursor ion is not readily apparent. Currently, there are a limited number of programs that exist to determine charge state from low resolution ETD MS/MS data, and only one of these is publicly accessible.\(^{34}\) This program, published by MacCoss and co-workers, takes advantage of a characteristic feature found in ETD spectra, the presence of intense peaks corresponding to charge reduced precursor species.\(^{34}\) However, this program has not been tested on precursor ions that contain disulfide linkages or more than one peptide. As it stands, all of the existing software programs that work to decipher charge state from ETD MS/MS data were designed for the analysis of peptides absent of intact disulfide bonds. The direct application of these automated tools to disulfide-bonded peptides is problematic in that peptides containing disulfide bonds have been shown to fragment differently when subjected to ETD in comparison to peptides without this covalent modification.\(^{12, 30}\) Therefore, a program specifically intended for charge state determination of peptides containing this common PTM is necessary.

In comparison to peptides, where the disulfide bonds have been reduced during sample preparation, peptides containing intact disulfide linkages show prominent peaks for both the
charge-reduced precursors and the individual peptides that comprise the disulfide-bonded precursor. Representative data from a lysozyme precursor with two peptides joined by one disulfide bond is shown below in Figure 1. In addition to the characteristic charge reduced precursor peaks, peaks corresponding to the individual peptides are also present in high abundance. As a result, programs that determine charge state by evaluating a mass spectrum for the relative abundance of those charge reduced precursors may not calculate charge state properly if the most prominent peaks detected instead correspond to individual peptide components.
Precursor charge state is not readily determined from low resolution ETD MS data, and no program is currently available to automate interpret ETD MS/MS data of disulfide-bonded precursors. To overcome these limitations in analysis, we developed a method that utilizes simple Excel-based tools to determine charge state from ETD spectra of intact disulfide-bonded peptides, and is capable of handling both types of native disulfide bonding arrangements. The premise of this method incorporates one accepted approach to deciphering charge state within a given mass spectrum; that is, by calculating the distance between adjacent peaks of the same compound that differ by a single charge (or one proton).

**Figure 1.** ETD MS/MS data collected at m/z 506.57 on a lysozyme precursor in the 3+ charge state with one interchain disulfide bond. Although peaks corresponding to the charge reduced precursor species are present in this spectrum, the most abundant product ions detected are the individual peptide chains that result from cleavage of the disulfide bond.

4.3.2 Method Development and Design. Precursor charge state is not readily determined from low resolution ETD MS data, and no program is currently available to automate interpret ETD MS/MS data of disulfide-bonded precursors. To overcome these limitations in analysis, we developed a method that utilizes simple Excel-based tools to determine charge state from ETD spectra of intact disulfide-bonded peptides, and is capable of handling both types of native disulfide bonding arrangements. The premise of this method incorporates one accepted approach to deciphering charge state within a given mass spectrum; that is, by calculating the distance between adjacent peaks of the same compound that differ by a single charge (or one proton). This is shown below by Equation 1.
Equation 1. Charge State \((z)\) of Peak 1 = \(\frac{\text{Peak 1}}{\text{Peak 2} - \text{Peak 1}} + 1\)

As charge reduced precursor ions are readily detectable within the ETD MS/MS data of disulfide-bonded peptides, MS/MS peak space information may then be used to determine the charge state of the precursor ion. If the charge state of the charged reduced precursor ion corresponding to one charge below that of the precursor ion is considered to be Peak 2, then the charge state of Peak 1 can be calculated by designating the precursor ion to be Peak 1.

However, there is no way to know which of the numerous peaks correspond to the charge reduced precursor species, and which correspond to other types of product ions generated, including those formed from each of the individual peptide chains when peptides with interchain disulfide bonds are present. Therefore, each of the peaks present in a tandem mass spectrum must be evaluated as potentially being Peak 2 and input separately into the charge state equation. After each of these independent calculations is performed, discriminatory analysis is necessary to determine which of the peaks corresponds to the actual charge reduced precursor species. Specifically, the charge reduced precursor one charge state below the parent ion. Using this premise, two simple and straightforward computational tools that work to automate the steps of this process were constructed in Excel.

To accomplish this, an ETD MS/MS peak list is first normalized to a 3 % relative abundance cut-off in order to reduce spectral noise. Relative abundance thresholds of 1 %, 2 %, 3 %, 6 % and 10 % were tested during method development. Next, \(m/z\) values for the product ions remaining in the normalized spectrum are imported into Excel. Here, two computational Excel-based tools work to determine the charge value associated with each of the MS/MS product ions, and then the actual charge state of the precursor ion.

In the first tool, the raw charge value for each peak present within an experimental ETD
spectrum, above the specified noise threshold, is calculated by evaluating the distance between the m/z values present in the MS/MS peak list and the m/z of the precursor ion. The function shown below by Equation 2 is input into an Excel spreadsheet to automate the charge value calculation associated each remaining product ion, where the m/z of the selected precursor ion is input into Column A and the MS/MS peak list data is input into Column B. The raw charge values are then output into Column C.

**Equation 2.** \( F_x (\text{Column C}) = \frac{\text{Column A}}{\text{Column B} - \text{Column A}} \)

Figure 2 provides an illustration of the first computational tool (constructed using Equation 2), as applied to a single ETD spectrum collected at m/z 818.03 from a lysozyme precursor.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Peak 1</td>
<td>Peak 2</td>
<td>Charge Value</td>
</tr>
<tr>
<td>2</td>
<td>818.03</td>
<td>1090.79</td>
<td>2.999</td>
</tr>
<tr>
<td>3</td>
<td>818.03</td>
<td>1091.54</td>
<td>2.991</td>
</tr>
<tr>
<td>4</td>
<td>818.03</td>
<td>1620.30</td>
<td>1.020</td>
</tr>
<tr>
<td>5</td>
<td>818.03</td>
<td>1622.43</td>
<td>1.017</td>
</tr>
<tr>
<td>6</td>
<td>818.03</td>
<td>1623.41</td>
<td>1.016</td>
</tr>
<tr>
<td>7</td>
<td>818.03</td>
<td>1627.38</td>
<td>1.011</td>
</tr>
<tr>
<td>8</td>
<td>818.03</td>
<td>1628.35</td>
<td>1.010</td>
</tr>
<tr>
<td>9</td>
<td>818.03</td>
<td>1629.19</td>
<td>1.008</td>
</tr>
<tr>
<td>10</td>
<td>818.03</td>
<td>1635.95</td>
<td>1.000</td>
</tr>
<tr>
<td>11</td>
<td>818.03</td>
<td>1637.65</td>
<td>0.998</td>
</tr>
<tr>
<td>12</td>
<td>818.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2.** Screen shot of computational tool 1 showing the Excel spreadsheet and charge value output for each peak in the normalized MS/MS peak list from a lysozyme precursor ion at m/z 818.03 in the 4 + charge state.

Using the second tool, discriminatory analysis is performed to choose the most probable
charge states based on constraints that were devised after extensive testing. Discerning potential charge values from all returned charge values is a two-step process. First, as the charge associated with a mass spectral peak must be an integer, the distance between each charge value and the nearest whole number is calculated. The following function shown by Equation 3 is input into Excel to accomplish this, where Column A is the raw charge value for each of the peaks in the MS/MS peak list (output from computational tool 1), and the distance between each individual charge value and the nearest integer is output in Column B.

\[ F_x (\text{Column B}) = \text{ABS}(\text{Column A} - \text{ROUND}(\text{Column A}, 0)) \]

In the second step, values that are beyond a specified range are eliminated and set to zero, and all charge values within a given error range are rounded to the nearest whole number. After extensive testing, the error threshold limit was set to 0.01, so any charge value that is greater than 0.01 away from an integer is eliminated. To automate this process, the function shown below by Equation 4 is input into Excel as Column C, and provides the final output of charge state for those remaining integers that fall within the acceptable error range. The integers that were eliminated are shown as 0. Finally, a custom sort function was used in Excel to sort the final charge states returned (Column C) on the basis of the integer rounding distance (Column B).

\[ F_x (\text{Column C}) = \text{IF(ABS(\text{Column B}))}>0.01,0,\text{ROUND}(\text{Column A},0) \]

As this method determines the charge state \( z \) of the charge reduced precursor ion that is one charge state below that of the precursor ion, the charge state of the precursor ion is equal to \( z + 1 \). The precursor charge state listed first in Column C of computational tool 2 is considered the most probable charge state for that charge reduced precursor species, except in the cases where
that number is one and another charge state of greater than 1+ is also returned. In these cases, if another integer greater than 1+ is listed, that integer is considered more probable. If two integers are given in the output list with equal or near equal rounding distances, a higher relative abundance threshold may be applied to the MS/MS peak list, as described later.

Figure 3 shows an example a typical charge state output for computational tool 2. In this run, the charge values were those integers computed for the MS/MS data shown by computation tool 1 in Figure 3.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charge Value</td>
<td>Rounding Distance</td>
<td>Charge State</td>
</tr>
<tr>
<td>2.999</td>
<td>0.0009</td>
<td>3</td>
</tr>
<tr>
<td>0.999</td>
<td>0.0010</td>
<td>1</td>
</tr>
<tr>
<td>0.998</td>
<td>0.0020</td>
<td>1</td>
</tr>
<tr>
<td>1.008</td>
<td>0.0085</td>
<td>1</td>
</tr>
<tr>
<td>2.991</td>
<td>0.0091</td>
<td>3</td>
</tr>
<tr>
<td>1.010</td>
<td>0.0095</td>
<td>1</td>
</tr>
<tr>
<td>1.011</td>
<td>0.0107</td>
<td>0</td>
</tr>
<tr>
<td>1.016</td>
<td>0.0157</td>
<td>0</td>
</tr>
<tr>
<td>1.017</td>
<td>0.0169</td>
<td>0</td>
</tr>
<tr>
<td>1.020</td>
<td>0.0195</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 3.** Screen shot depicting computational tool 2. The charge values shown in Column A are the output from computational tool 2. For the MS/MS data shown here, a charge state 3+ for the charge reduced precursor species in the precursor – 1 charge state, the disulfide bonded precursor is a 4+ charge state ion.

### 4.3.3 Precursor Charge State Assignment of Disulfide ETD MS/MS Data

Disulfide-bonded precursors in various charge states from lysozyme, fetuin, BSA, and transferrin were assigned charge state using the two computational tools from the method described in the previous section.

Representative data from a fetuin precursor in the 4+ charge state is shown below in
Figure 4. In this example, the total combined length of the two peptides comprising the precursor ion is 27 amino acids. After the MS/MS peak list from an ETD spectrum collected at $m/z$ 724.34 was tested, the precursor ion was determined to be in the 4+ charge state. This assignment is correct, according to manual verification. In Figure 4, peaks corresponding to individual chains of the disulfide bonded precursor are present in high abundance. This agrees with previous research indicating the preferential cleavage of disulfide bonds during ETD.\textsuperscript{10,12,13}

![Figure 4. ETD MS/MS data at $m/z$ 724.34 collected on fetuin. The fragmentation shown by the spectrum is representative of a precursor ion with intact interchain disulfide bonding. Product ions resulting from the cleavage of the disulfide bond, and the characteristic charge reduced precursors, are both present in high abundance.](image)

One of the initial difficulties in designing a method for precursor charge state
determination arose from the differences in the intensity of the charge reduced precursor peaks that were observed for disulfide bonded precursors of different charge states. This coincides with the different degrees of accuracy that MacCoss and co-workers reported for precursor ions of different charge state in their peptide ETD MS/MS program.\textsuperscript{34} The reported accuracy was over 99\% for precursors of 2+, 3+, and 4+ charge states, but significantly lower for spectra collected on peptide precursors with a higher charge state.\textsuperscript{34} Therefore, it was important to ensure that the normalization level applied to the MS/MS peak list be applicable to all disulfide bonded precursors, regardless of charge state. An example of a spectrum collected on a 5+ charge state transferrin precursor ion and scored using the devised charge state analysis method is shown in Figure 5. The spectrum was analyzed using computational tools 1 and 2, and the precursor was correctly assigned as a 5+ charge state ion.
Often, more than two individual chains are joined to form a precursor containing native disulfide bonding. Therefore, it was important to ensure that our charge state determination tools were applicable to these disulfide-bonded peptides as well. The ETD spectrum collected on a precursor from BSA that contains two disulfide bonds and three peptide chains is shown below in Figure 6. In this case, a 5+ charge state assignment after the MS/MS data was evaluated using the computational tools described herein.

Figure 5. ETD MS/MS data collected at m/z 565.67 on a transferrin precursor comprised of two peptides joined together by one interchain disulfide bond. In agreement with the tests of ETD spectra obtained on disulfide bonded peptides in the 2+, 3+, and 4+ charge states, our computational approach also identified the charge state for 5+ precursor ion.
Further adding complexity to the ETD MS/MS analysis of disulfide bonded peptides are the differences reported for fragmentation reported for interchain and intrachain bonds types. These include variations seen in both the amount, and type, of product ions formed. These differences have also been observed for experimental ETD data acquired in our lab. For these reasons, it was important to develop a method for determining charge state that was applicable to both types of disulfide bond arrangements. Experimental ETD MS/MS data of a transferrin peptide containing three intrachain disulfide bonds is depicted in Figure 7. Using the same computational tools, this precursor was correctly identified as a 4 + charge state ion.

Figure 6. ETD MS/MS data collected on bovine serum albumin at m/z 770.79, in the 5 + charge state. This precursor ion consists of three peptides joined together by two disulfide bonds.
In total, over 70 ETD spectra from lysozyme, fetuin, BSA, and transferrin were assigned charge state using the Excel-based computational tools developed in our lab. MS/MS data from precursors of various charge state and bonding arrangements were tested using this method, with an accuracy of over 90% for all spectra combined. A future direction of this project is to apply the computational tools to a validation set of spectra from a different protein.

4.3.4 Number of Charge State Assignments Returned. One of the limitations to current programs designed for charge state determination of peptide ETD MS/MS data is the inability to assign a single charge state to a precursor ion. The number of charge states returned to a user per spectrum tested is referred to as the z: scan ratio, and attaining the lowest value
possible is reportedly a high priority goal for other programs aiming to decipher precursor charge state.\textsuperscript{34} In short, the $z$: scan ratio of 1.00 is advantageous for researchers working to limit analysis to the shortest time possible.

For example, the recommended user parameter for the previously mentioned SVM classifier tool is a cut-off probability of 0.98, which corresponds to a $z$: scan ratio of 1.53, as reported for the LTQ-ETD unique peptide data set.\textsuperscript{34} This means that the meaning SVM classifier will aim to generate 1.53 charge state predictions for each spectral search performed.\textsuperscript{34} So, out of the six possible charge states considered by the tool (2+ through 7+), close to two charge state predictions are returned to a user for each query.\textsuperscript{34} Each additional charge state assignment significantly increases analysis time for researchers using database searches to identify peptide MS/MS data, as each one requires a separate database query to identify the precursor ion using automated search algorithms. One the other hand, accuracy typically improves with a higher $z$: scan ratio, as all potential charge states are returned to a user for a given spectrum. Therefore, using the default parameters of this program for the described data, the program achieves an accuracy of almost 99\%.\textsuperscript{34} This is significantly higher than the accuracy reported for charge state assignment when a $z$: scan ratio of 1.00 is selected for the same data, which was below 97\%.\textsuperscript{34}

To overcome this limitation for researchers utilizing a database capable of processing peptide ETD spectra with intact disulfides, we devised a computational method that predicts the most probable charge state for a precursor ion. As such, a charge state prediction return of 1.00 can be achieved for each spectrum tested. Other potential charge states are also returned to the user, in order of decreasing probability. This is beneficial for users who are more interested in accuracy than speed of analysis.
As previously described, the top precursor charge state returned to a user is considered the most probable charge state for that charge reduced precursor species, except in the cases where that number is one. For these, if a second integer greater than one is also returned, that integer is considered more probable. In cases where two integers are given in the output list with equal or near equal rounding distances, it is suggested that a higher relative abundance threshold of 10% is applied to the MS/MS peak list, and the process repeated. Figure 8 illustrates an ETD spectrum of a 4+ transferrin precursor containing two disulfides, and shows typical product ions that result from the cleavage of interchain bonds. When this spectrum was initially analyzed using our method, two different charge states were returned when the standard 3% relative abundance cut-off was applied. When the normalization was increased to 10%, a precursor charge state of returned. This assignment is in agreement with the manual assignment.
Although a few programs exist to aid in the interpretation of low resolution peptide ETD spectra, no method has previously been described for the assignment of disulfide-bonded peptide ETD spectra. To overcome this need, we have created an algorithm that allows the determination of precursor charge state directly from low resolution ETD MS/MS data.

This simple approach utilizes simple computational tools to allow a user to quickly access the most likely charge state directly from experimental MS/MS data, bypassing the need

**Figure 8.** ETD MS/MS data at m/z 941.48 collected on transferrin. A total of two interchain disulfide bonds join the three peptide chains of this 4+ charge state precursor ion. Individual chains of the disulfide bonded precursor ion were readily detectable in the fragmentation profiles of precursors containing interchain disulfide bonds.

**4.4 CONCLUDING REMARKS**

The computational method presented herein is designed to overcome the most significant limitations of current tools that work to determine precursor charge state from ETD MS/MS data. Although a few programs exist to aid in the interpretation of low resolution peptide ETD spectra, no method has previously been described for the assignment of disulfide-bonded peptide ETD spectra. To overcome this need, we have created an algorithm that allows the determination of precursor charge state directly from low resolution ETD MS/MS data.

This simple approach utilizes simple computational tools to allow a user to quickly access the most likely charge state directly from experimental MS/MS data, bypassing the need
to rely on isotopic distribution patterns for deciphering charge state of disulfide-bonded peptides.
In addition, this method is advantageous in that no additional computer downloads are needed, no Linux operating system is necessary, and no learning curve is required before use.

4.5 ACKNOWLEDGEMENTS

The author acknowledges financial support from an NSF CAREER Award (0645120) to H.D., an NSF Fellowship (DGE-0742523) and Pfizer Scholarship to C.W., and a Seo Scholarship to M.M.

The author also wishes to thank those who contributed to the work described herein: Daniel Clark for providing some of the ETD MS/MS data, Morgan Maxon for her time and contribution, and Heather Desaire for her time and leadership.
4.6 REFERENCES


CHAPTER 5

FUTURE DIRECTION: GLYCOPEP GRADER UPDATES

ABSTRACT

GlycoPep Grader (GPG) is a publicly available tandem mass spectrometry (MS/MS) data analysis tool that scores glycopeptide candidate compositions by evaluating two types of product ions: 1) Ions that contain the peptide plus some portion of the pentasaccharide core, referred to as [peptide + core component] ions, and 2) Ions formed via the neutral loss of monosaccharide residues from the precursor ion, or [precursor – monosaccharide] ions. Although GPG has shown unprecedented success in identifying the correct glycopeptide candidate composition for a given CID spectrum, a number of vital updates that should work to create a larger separation in scores among the correct and incorrect compositional assignments have been identified.

Specifically, there are adjustments in the scoring algorithm that could be made for the [peptide + core component] product ions, which are applicable to all N-linked glycopeptide types, as well as a number of changes for glycopeptides containing complex or hybrid type glycans. The new rules proposed for grading the [precursor – monosaccharide] product ions for these species would affect both the glycopeptide marker ions and precursor neutral losses currently searched by GPG. These results are so far untested and based on observations made for the original collection of CID data that was used train and validate the software. Most notable are suggested improvements to account for complex and hybrid type glycopeptide arrangements that contain sialic acid residues.
5.1 INTRODUCTION

The interpretation of glycopeptide data from tandem mass spectrometry (MS/MS) experiments remains challenging today, even with the advent of automated tools to assist in the elucidation of these spectra. When glycopeptides containing sialic acid are considered, glycopeptide identification tends to become even more difficult, due to the inherent problems associated with the characterization of these negatively charged residues when using positive ion mode collision induced dissociation (CID).\textsuperscript{1,2,3,4} Although the negative charges add complexity to their analysis, these acidic glycans play major functional roles in biological processes.\textsuperscript{4,5,6,7,8} In addition, sialic acids have proven critical to the development of efficacious and safe glycoprotein therapeutics.\textsuperscript{6,9,10,11,12,13} For example, the presence of sialic acid residues in complex or hybrid glycans has been shown to increase the circulatory half-life of erythropoietin in comparison to the asialo erythropoietin counterpart.\textsuperscript{6,9,10,11,12}

To date, one of the current restrictions to publicly available glycopeptide software is the inability to accommodate the distinctive CID features imparted by glycopeptides containing sialic acid. Although the MS/MS data collected on glycopeptides of high mannose or complex/hybrid asialo type glycans generally contain more identifying fragmentation features than those containing sialic acid, most automated programs fall short in their analysis as well. One of the ways to improve the automation of glycopeptide identification is to make use of CID product ion intensity information, which limited algorithms are equipped to do.\textsuperscript{14} More challenging still, recent studies have shown that the charge state of a precursor ion is a critical component to accurately identifying many of the product ions expected to be present upon fragmentation of a given glycopeptide.\textsuperscript{14,15} This applies to all glycopeptides, though it is amplified for sialylated species.
Few of the currently available automated MS/MS data analysis programs were specifically intended for the characterization of glycopeptide spectra; they were instead, designed for glycans. Accordingly, these programs lack the capacity to investigate the fragmentation profiles of both the peptide and glycan portions of a glycopeptide.\footnote{15} This is problematic in glycopeptide analysis because many arrangements of these two unknowns (peptide and glycan portions) combine to form a nearly identical neutral mass.\footnote{16} It was with the goal of overcoming these debilitating limitations that GlycoPep Grader (GPG) was originally constructed. GPG’s unique capabilities ultimately allow a user to discriminate between isobaric $N$-linked glycopeptide compositions. The correct composition is determined by scoring experimental MS/MS data in a highly specific manner that depends on the fragmentation patterns typical of each type of glycan substituent attached to a peptide.\footnote{15}

Although GPG has shown unprecedented success in the identification of $N$-linked glycopeptide compositions, a number of potential improvements to the program have been identified. The most important of which should greatly improve the scoring of sialylated glycopeptides. These proposed changes are based on the recent discovery of additional fragmentation patterns found to be present in CID spectra of $N$-linked glycopeptides containing sialic acid. In addition, the detailed findings described herein are not only essential for the enhancement of GPG, but are also significant because they lend further insight into charge state dependent fragmentation of glycopeptides, which has not been studied in much detail.\footnote{17,18}

Additional updates to GPG are also discussed below. One of these comes from an important observation made for all MS/MS data collected on $N$-linked glycopeptides containing hybrid or complex type glycans. Specifically, a change in a typical glycopeptide marker ion, or commonly observed oxonium ion, is prosed for future versions of GPG. The oxonium ion is
significant for the software because the currently used marker ion, at m/z 366, is often out of the scan range, and therefore not evaluated by GPG for a majority of CID spectra. Herein, we change this marker ion to m/z 528, which is more likely to be detected because it has a greater probability of being within the MS/MS scan range.

Overall, the potential improvements that have been identified since the release of GPG should decrease variations in scoring, especially in the case where complex glycopeptides are modified by sialic acid. To this end, these improvements should work to increase user confidence in the results for those tests where scores between alternate compositional assignments are not as pronounced as in most cases.

5.2 EXPERIMENTAL

5.2.1 Materials and Reagents. Bovine asialofetuin, bovine ribonuclease B (RNase B), human apo-transferrin (transferrin), urea, dithiothreitol (DTT), iodoacetamide (IAM), formic acid, acetic acid, Sepharose® CL-4B, HPLC grade ethanol, and HPLC grade 1-butanol were purchased from Sigma Aldrich (St. Louis, MO). HPLC grade methanol (CH₃OH) and HPLC grade acetonitrile (CH₃CN) were purchased from Fisher Scientific (Fairlawn, NJ). Ammonium bicarbonate (NH₄HCO₃) was purchased from Fluka (Milwaukee, WI) and sequencing grade modified trypsin was from purchased Promega (Madison, WI). Ultrapure water was obtained from a Millipore Direct-Q® UV 3 system (Billerica, MA) with a resistance greater than 18 MΩ.

5.2.2 CID MS/MS Data of RNase B, Asialofetuin, and Transferrin Glycopeptides. Detailed information on the preparation and MS analysis of RNase B, asialofetuin, and transferrin samples can be found in Chapter 2 of this dissertation. Briefly, samples of RNase B, asialofetuin, and transferrin were each prepared by Rebecchi and Woodin on multiple occasions, using an enrichment method for the RNase B glycopeptides that was developed by Rebecchi et
This collection of CID spectra, obtained for each sample by Rebecchi and Woodin, results from the compilation of experiments published in the original GPG article, therein referred to as the glycopeptide training data set.\textsuperscript{15}

5.2.3 CON-S gp140 CFI Preparation and CID MS/MS Data. The CON-S gp140 CFI CID spectra shown herein are from the glycopeptide validation data set of the original GPG article, which was originally analyzed and reported on by Go et al.\textsuperscript{15,20} Sample preparation for the CON-S gp140 CFI glycopeptides was also performed by Go and co-workers, as previously described.\textsuperscript{20} Detailed information on the experimental procedures and MS analysis is also given in Chapter 3 of this dissertation.

5.2.4 Manual Data Analysis. To identify the glycopeptides from these samples in the MS data, a prediction table of theoretical $m/z$ values corresponding to glycopeptide compositions for RNase B, asialofetuin, and transferrin was prepared. The amino acid sequences from the proteins were obtained from Uniprot (www.uniprot.org) and their sequences were imported into Protein Prospector (http://prospector.ucsf.edu/prospector/mshome.htm). In the Protein Prospector programs, settings were used to indicate peptides containing Cys residues were modified with carboxamidomethylation, and a theoretical tryptic digest was performed to consider up to two tryptic miscleavages. Peptide masses containing potential N-linked glycosylation sites were added to the mass values of known, biologically relevant glycans, in order to obtain glycopeptide masses. These theoretical glycopeptide masses were converted into $m/z$ values corresponding to the glycopeptides existing in multiple charge states. The MS/MS data for RNase B, asialofetuin, and transferrin were then searched to identify spectra that contained $m/z$ values for ion fragments that correspond to the theoretical $m/z$ values for a given glycopeptide composition. The CID spectra were carefully (manually) evaluated in order to verify the
glycopeptide assignment, and subjected to analysis by GPG.

5.3 RESULTS AND DISCUSSION

The novel GPG software developed in our lab is a great advancement for researchers working to decipher glycopeptide composition from MS/MS data. Although they have not yet been tested, changes to the original GPG algorithm that could improve scoring for all N-linked glycopeptides have been identified. A summary of the major product ions for each glycopeptide type is shown in Figure 1, as a reminder. This figure is also shown in Chapter 2 of this dissertation, where the basis for these devised glycan categories and their illustrated product ions is described in great detail.
A. [Peptide + Core Component] Ions

B. [Precursor – Monosaccharide] Ions

Figure 1. Schematic of (A) [peptide + core component] and (B) [precursor – monosaccharide] product ions expected for each of the eight group types, described in the text. In (A), the six different [peptide + core component] product ions detected for a glycopeptide, regardless of glycan type, are displayed. In (B), the monosaccharide neutral losses evaluated for group 1 are shown in the purple oval; for group 2, the relevant losses are shown in both the purple oval and the yellow circle; group 3, the relevant losses are shown in the blue oval; group 4, in the blue oval and yellow circle; group 5, in the orange oval and yellow circle; group 6, in the green oval and yellow circle; and group 7 and group 8 neutral losses are presented by the orange oval, and the green oval, respectively. This figure is adapted from the original ACS publication on GPG.15

The impact of the scoring differences is expected to be dependent upon the monosaccharide arrangements comprising the appended glycan, which the most pronounced differences expected for complex and hybrid type glycans, especially for those containing sialic acid. This is hypothesized because there is not always a large difference in GPG scores between the actual and decoy candidate compositions for these compositions, as evidenced by some test cases shown in Chapter 3, Tables 2 and 3, of this dissertation. For a few specific examples, see Tests 27, 28, and 29 of Chapter 3, Table 2. Other updates should improve scoring for all N-linked glycopeptides regardless of glycan substituent, such as those pertaining to the [peptide + core component], or peptide-containing, product ions. The proposed updates that have been
formulated, but not yet tested, to improve scoring in future GPG software versions are described below.

5.3.1 Peptide-Containing Glycopeptide Product Ions. In order to minimize the contribution of random peak matches during the MS/MS data search by GPG, the following update to the GPG algorithm is proposed. Currently, each match identified within the MS/MS peak list that corresponds to those [peptide + core component] ions expected to be present for a given glycopeptide is assigned a uniform score. Instead, the algorithm could be updated to incorporate an additional scoring factor for each consecutive hit by GPG for these ions. This update would increase the TotalRawPeptideScore (theoretical points possible) associated with each composition. Then, when each of these respective product ions found in series after the first one is detected, an incrementally higher point value would be assigned to each candidate’s respective ActualRawPeptideScore (actual points awarded). The implementation of this scoring factor would affect the overall PeptideScore (ActualRawPeptideScore/TotalRawPeptideScore) associated with each candidate glycopeptide composition, though the final GlycopeptideScore would still be weighed the same. That is, the PeptideScore would still account for 67% of the total GlycopeptideScore. The way the scoring of these terms are calculated is detailed in the original GPG algorithm, as shown in Chapter 2, Table 2, of this dissertation.

Essentially, the longest number of consecutive [peptide + core component] product ions that could be detected for each candidate composition would be added to their respective TotalRawPeptideScore, and the number of these found in series would be added to their ActualRawPeptideScore. This would be done for both the charge state of the precursor ion and for the charge state below the precursor ion, as long as the calculated product ions are within scan range. For example, if matches were detected for a candidate composition that
corresponded to the [peptide + HexNAc], [peptide + 2HexNAc], [peptide + 2HexNAc + Hex], and [peptide + 2 HexNAc + 2Hex] ions, but no product ions was detected for the calculated [peptide] or [peptide + 2 HexNAc + 3Hex] ions, the longest number of consecutive matches for the peptide-containing product ions would be 4. Therefore, 4 points would be added to that glycopeptide’s \textit{ActualRawPeptideScore}. However, since there were a total of 6 peptide-containing product ions possible, 6 points would be added to that candidate’s \textit{TotalRawPeptideScore}. The purpose of this would be to decrease the impact of random peak matches, especially in the case of spectra containing high noise levels.

5.3.2 Sialylated Glycopeptides. Recent studies suggest that fragmentation of glycopeptides containing sialic acid is profoundly impacted by precursor charge state, and that more than one dissociation pathway may to take place for these species under different charge states.\textsuperscript{14, 15, 17, 21} Although a charge that is imparted during the electrospray ionization process may reside in one of two distinct locations within the glycopeptide, namely the peptide backbone or the attached glycan moiety, it is assumed that a glycan will not support more than one charge due to a lack of basic (proton accepting) sites and inherent Coulomb repulsion between charges.\textsuperscript{14} However, glycosidic cleavages may result from two distinct mechanisms, charge-remote pathway or charge-directed pathway.\textsuperscript{14, 17} Proton distribution is shown to affect the probability of each pathway, especially in the case of sialylated glycopeptide precursors.\textsuperscript{14} For these species, it has recently been shown that precursor charge state determines which of the cleavage pathways will dominate.\textsuperscript{14}

Through extensive CID MS/MS data analysis performed in our lab, it was demonstrated that fragmentation patterns in sialylated glycopeptides are dependent on charge state. For precursor ions with a charge state \textit{greater than or equal to} 3 +, a different fragmentation profile
was observed in the experimental data of those \( N \)-linked glycopeptides containing sialic acid, in comparison to those precursors in the 1 + and 2 + charge states. These two fragmentation patterns are depicted below in Figure 2. In A, the most likely point of cleavage is the glycosidic bond joining sialic acid, or Neu5Ac, to the rest of the glycan substituent. This leads to the loss of individual sialic acid residues from the glycopeptide precursor. In B, the most likely point of cleavage is the glycosidic bond after the HexNAc residue of the glycan antennae, which results in the loss of the entire branch, or the combined residues of Neu5Ac + Hex + HexNAc from the precursor ion.

**Figure 2.** Illustration of the two different fragmentation pathways observed in CID spectra of sialylated glycopeptides. In (A) the predominant cleavage observed for glycopeptide precursors in the 1 + and 2 + charge states is shown, whereas the most likely point of cleavage detected for glycopeptides in charge states higher than 3 + is shown in (B). According to experimental CID spectra, glycopeptide precursor ions with a charge state of 3 + were found to frequently dissociate by either pathway.

The experimental data obtained in our lab corroborates the mathematical predictions of Zhang, who recently extended a peptide fragmentation model to experimental CID MS/MS data of \( N \)-linked glycopeptides.\(^ {14, 22, 23} \) In these studies, Zhang and Shah demonstrate that sialylated glycopeptides will dissociate according to charge-remote or charge-directed cleavage depending on charge state, with characteristic losses for each matching the fragmentation pathway shown in Figure 2A and Figure 2B, respectively.\(^ {14} \) However, they report that sialic acid residues with a
charge state equal to 3+ were found to dissociate by the charge-remote cleavage indicative of lower charge state species. In comparison, according to our collection of CID MS/MS data, these 3+ glycopeptide species dissociate by both the charge-remote and charge-directed pathways.

To illustrate the charge-directed cleavage of a 3+ precursor ion containing sialic acid, MS/MS data of a transferrin glycopeptide is shown in Figure 3. Here, the predominant neutral losses detected within a CID spectrum are different than those expected for sialylated glycopeptides of lower charge state. Specifically, a peak corresponding to the loss of an entire branch, or the combined residues of HexNAc + Hex + Neu5Ac from the precursor ion, was found to be present in high abundance for those species of higher charge states.
Figure 3. CID data collected on a transferrin glycopeptide at \( m/z \) 1227.9 in the 3 + charge state. The most abundant product ion in this spectrum is detected in the 2 + charge state, resulting from a combined loss of HexNAc + Hex + Neu5Ac residues from the precursor. A very intense peak at \( m/z \) 657 is also present in the spectrum, and is indicative of these species. Although the losses of individual sialic acid residues are also present within the MS/MS data, these product ions are much less abundant than they are for sialylated glycopeptide precursors of charge states lower than 3 +. The current version of GPG scores the loss of individual sialic residues, and does so at a relative abundance threshold that is above the detection limit for those charge states greater than or equal to 3 +.

In contrast, the most predominant neutral fragment for a precursor ion of lower charge state species was found to result from a loss of individual sialic acid residues, in both the charge state of the precursor and the charge state below the precursor ion. These losses are illustrated in Figure 4. Evaluation of glycopeptide candidate compositions in future versions of GPG should benefit by incorporating both fragmentation pathways into the scoring algorithm.
Figure 4. MS/MS data at m/z 1841.8 of a sialylated glycopeptide from transferrin in the 2+ charge state. For the CID data shown here, the most product ions are formed from individual losses of sialic acid from the precursor, which are detected in the 2+ charge state. A peak at m/z 657 is present in the spectrum as well, but is less prominent than the same marker ion depicted in Figure 3.

The spectra of the higher charge state species are also shown to contain an intense glycopeptide marker ion at m/z 657, which is also depicted in Figure 3. This marker ion is present in the spectra of most compositions that contain sialic acid, regardless of charge state. However, the relative abundance of the peak may vary. Another proposed update to GPG scoring is to score the presence of this marker ion for all sialylated glycopeptides, regardless of precursor charge state.

5.3.3 Glycopeptide Marker Ion Detection for Complex/Hybrid Type Glycans. The algorithm behind GPG could be improved if an update was made to score the glycopeptide
marker ion of $m/z$ 528, as opposed to the glycopeptide marker ion of $m/z$ 366. One of the reasons for this is ambiguity of the $m/z$ 366 ion that arises when scoring complex/hybrid type glycopeptides against CID spectra of high mannose type glycopeptides, as the MS/MS data collected on these species generally contains the $m/z$ 366 marker ion as well. This trend is also shown by Huddleston et al. who concluded that CID data of glycopeptides with high mannose and complex branching give rise to an intense peak at this molecular weight, due to the loss of HexNAc + Hex from the charged precursor ion.\textsuperscript{27} Furthermore, from the experimental MS/MS data acquired in our lab, glycopeptides containing sialic acid were also shown to possess a peak at $m/z$ 366. The presence of this marker ion in the CID spectra of sialylated glycopeptides has been also reported by Conboy and Henion.\textsuperscript{28}

Often, because of the low molecular weight scan cut-off used, the marker ion at $m/z$ 366 is out of scan range when glycopeptide MS/MS data is acquired on tryptic digests of glycoproteins. This occurs because the lowest mass range is limited on an ion trap instrument to approximately 1/3 of the precursor ion’s $m/z$.\textsuperscript{29} GPG currently searches for the loss of this specific product ion for glycopeptide compositions that contain more terminal Hex than HexNAc residues, as shown in Figure 5. Therefore, in many instances, the glycopeptide marker ion’s score is not factored into the overall candidate score. This adversely affects scoring for a number of compositions.
Figure 5. Schematic of the [precursor – monosaccharide] product ions searched by GPG for each of the eight glycopeptide group types, previously detailed in Chapter 2 of this dissertation. The glycopeptide marker ion at \( m/z \) 366 is currently searched for complex and hybrid type glycans belonging to group 6 (which include those glycan compositions that contain at least one fucose, and more terminal Hex than HexNAc residues) and group 8 (which include those glycan compositions with no appended fucose, and more terminal Hex than HexNAc residues). The characteristic product ions for group 6 are depicted by the yellow circle and green oval, whereas the characteristic product ions for group 8 are shown in the green oval. This figure is adapted from the original ACS publication on GPG.^^15^^

In Figure 6, representative MS/MS data collected on an asialofetuin glycopeptide with a scan range of 500-2000 \( m/z \) is given. The composition of the glycan in this case contains more terminal Hex than HexNAc residues. Although \( m/z \) 366 is out of range, the spectrum shows the presence of a peak at \( m/z \) 528, which corresponds to the oxonium ion that results by the loss of Hex + HexNAc + Hex from a complex type glycopeptide precursor.
Figure 6. CID spectrum from a complex type asialofetuin N-linked glycopeptide precursor in the 2+ charge state. An intense marker ion at $m/z$ 528 is shown to be present in the spectrum. The current version of GPG evaluates the presence of a marker ion at $m/z$ 366. This product ion is often out of scan range for MS/MS data taken on tryptically digested glycopeptides, whereas $m/z$ 528 is more likely to be within the scan range.

Originally, the algorithm was designed to search for a peak at $m/z$ 366 for only those complex and hybrid type glycans, absent of sialic acid residues, that contain an appropriate number of Hex versus HexNAc residues. The expected fragmentation for glycopeptides of these glycan categories is shown by groups 6 and 8 of Figure 1. The characteristic [precursor – monosaccharide] product ions for these glycan compositions is shown again in Figure 5, although a more thorough discussion on each of the devised glycan categories (including their respective monosaccharide arrangements) is provided in Chapter 2 of this dissertation. After extensive analysis of experimental CID MS/MS data, it became apparent that both of these
marker ions (m/z 366 and m/z 528) are detected in the spectra of most complex and hybrid N-linked compositions, regardless of the ratio of terminal residues. Therefore, future versions of GPG should score the presence of this ion for all complex and hybrid type arrangements that do not contain sialic acid, regardless of whether they contain more HexNAc or Hex residues after the common pentasaccharide core. The original fragmentation evaluated for these glycans, which this new rules will now extend to, is illustrated in groups 5 and 7 of Figure 1.

In CID experiments fucose is considered labile, and numerous instances of rearrangement events have also been documented for fucosylated species.\textsuperscript{17, 30} For complex/hybrid type glycopeptides appended with fucose, GPG not only evaluates the presence or absence of the product ions formed by the loss of this monosaccharide, but the remaining glycan portion as well. As such, it was important to verify the formation of the oxonium ion at m/z 528 for fucosylated complex type glycopeptides to ensure that this update would be applicable to these more labile compositions. In Figure 7, the presence of this marker ion is shown to be readily detectable for an N-linked glycopeptide from CON-S gp140 CFI containing one fucose residue. In contrast to the data shown in Fig. 6, the glycan substituent in this instance contains more terminal HexNAc than Hex residues.
Figure 7. MS/MS data collected on a fucosylated N-linked glycopeptide from CON-S gp140 CFI. An intense marker ion at m/z 528 is shown to be present for complex/hybrid compositions whether or not fucose is present. This is significant because GPG evaluates the remaining portion of a glycopeptide independently after the expected product ions resulting from loss of fucose are scored.

Due to the issues discussed above, the glycopeptide marker ion at m/z 528 is presumed to be a more logical choice for the identification of complex or hybrid type glycopeptides over the marker ion at m/z 366, which is currently evaluated by GPG when scoring appropriate [precursor – monosaccharide] product ions. A schematic of future GPG scoring encompassing all of the proposed updates to the [precursor – monosaccharide] product ions, including the changes in glycopeptide marker ion evaluation, is given in Figure 8.
Figure 8. Proposed schematic of [precursor – monosaccharide] product ions searched by future versions of GPG for scoring each of the eight glycopeptide group types. The monosaccharide neutral losses evaluated for group 1 are shown in the purple oval; for group 2, the relevant losses are shown in both the purple oval and the yellow circle; group 3, the relevant losses are shown in the blue oval; group 4, in the blue oval and yellow circle; group 5, in the orange oval and yellow circle; group 6, in the green oval and yellow circle; and group 7 and group 8 neutral losses are presented by the orange oval, and the green oval, respectively. Expected product ions for group 1 and group 2 remain unchanged if these updates are incorporated into the current algorithm, but change for complex/type glycans with (groups 3 and 4) and without (groups 5, 6, 7, 8) sialic acid. This figure is adapted from the original ACS publication on GPG.\textsuperscript{15}

5.3.4 Other Potential Future Updates. One of the more obvious improvements to future versions of GPG is the incorporation of a scoring function designed to handle glycopeptide precursors containing common adducts, such as sodium or potassium. These adducts are often found in CID MS/MS data of N-linked glycopeptides, especially those samples extracted from biological matrices.\textsuperscript{21,31,32} However, other updates have been identified as more vital at this time. In addition, as the development of GPG is complex and heavily based in
mathematical calculations, the incorporation of scoring designed for glycopeptide adducts would take a long time to implement.

Another future update to GPG centers on analysis of glycopeptide branching characteristics. Although the software only identifies composition at this point, future versions may allow users the ability to decipher basic structural information such as the number of glycan branches. One of the recent observations made in our laboratory for CID spectra of complex or hybrid type glycans is a difference in intensity for two common product ions. Specifically, a comparison in the intensity of product ions formed by loss of 1) HexNAc, and 2) HexNAc + Hex, from the glycopeptide precursor. A comparison between the intensity of these two product ions may lend potential insight into glycopeptide branching, including whether a glycoform is bi- or tri-antennary.

Finally, a potential future direction of GPG could be to incorporate an algorithm into the software that is specific to O-linked glycopeptides. Currently, no automated analysis program has been designed to evaluate MS/MS data and determine the identity of both the peptide and glycan portions of O-linked glycopeptides. Some of the challenges associated with the automation of O-linked glycopeptide MS/MS data analysis are given in Chapter 1 of this Dissertation. The inclusion of a GPG algorithm designed specifically for MS/MS scoring of O-linked species would be a major project, as a new set of fragmentation rules would first need to be devised using CID MS/MS data collected on O-linked glycopeptides.

5.4 CONCLUDING REMARKS

Recently, we developed an automated analysis tool, GlycoPep Grader (GPG) to determine the N-linked glycopeptide composition for a given CID spectrum. After extensive testing of the program, a number of improvements that could render the GPG more effective
have been identified. In addition, although the GPG program has shown to be highly accurate, a limited number of spectra collected on glycopeptides containing sialic acid, or both sialic acid and fucose, have been evaluated for improved data analysis.

When testing MS/MS data of N-linked glycopeptides containing these labile modifications, the scores between actual and decoy glycopeptide compositions of nearly identical neutral mass were found to be closer in value than the scores acquired for glycopeptide precursors possessing other N-glycan arrangements, as shown in Tables 2 and 3 from Chapter 3 of this dissertation. After further analysis, an alternate predominant fragmentation pathway was identified for glycopeptide precursor ions containing sialic acid appearing in the 3 + charge state or higher. This pathway demonstrates the consistent loss of a sialylated glycan branch. The incorporation of this pathway, along with other described improvements to the algorithm, should improve the scoring margin of GPG when discriminating the actual glycopeptide composition from a pool of decoy candidates corresponding to the same m/z.

5.5 ACKNOWLEDGEMENTS

The author acknowledges financial support from the NIH (RO1RR026061) to H.D., and an NSF Fellowship (DGE-0742523) and Pfizer Scholarship to C.W.

The author also wishes to thank those who contributed to the work described herein: David Hua for writing the final version of the algorithm, Morgan Maxon for her contribution to the development and testing of GPG, Eden Go for supplying the CON-S gp140 CFI MS/MS data, Katie Rebecchi for providing some of the glycopeptide CID spectra collected on transferrin, and Heather Desaire for her suggestions and advice.
5.6 REFERENCES


(27) Huddleston, M. J.; Bean, M. F.; Carr, S. A. Collisional fragmentation of glycopeptides by


CHAPTER 6

CONCLUSION

6.1 SUMMARY OF DISSERTATION CONTENT.

Protein glycosylation and disulfide bond formation are two common post-translational modifications (PTMs) that are widespread among all three taxonomic domains.\(^1\),\(^2\),\(^3\) The characterization of these and other PTMs are routinely performed using mass spectrometry (MS) and tandem mass spectrometry (MS/MS) experiments,\(^4\),\(^5\) and the availability of reliable automated tools greatly increases the amount of data that can be processed in a given amount of time.\(^6\)

In the analysis of glycopeptides, a lack of publicly available programs to evaluate the two individual components they are comprised of, the peptide and glycan portions, has hampered the speed at which collision induced dissociation (CID) MS/MS data interpretation is accomplished.\(^6\),\(^7\) To overcome this limitation, we developed the GlycoPep Grader (GPG) program to automate the compositional determination of \(N\)-linked glycopeptides from CID spectra.\(^7\) The algorithm that powers GPG was designed using a set of glycopeptide fragmentation rules derived from careful analysis of experimental CID MS/MS data collected on glycopeptides with various \(N\)-linked glycan arrangements.\(^7\)

The ability to map a protein or peptide’s disulfide bonds through MS/MS experiments is plagued with challenges similar to those encountered in the elucidation of glycopeptide MS/MS data. That is, automation of applicable programs to process MS data of peptides containing disulfide bonds is still in the infancy stage. Furthermore, software for newly developed dissociation techniques such as electron transfer dissociation (ETD) MS/MS were not designed to interpret data collected on proteins or peptides with intact disulfide linkages.\(^8\) As a result,
analysis tools to assist in the extraction of even basic information from a mass spectrum, such as precursor charge state, are currently lacking for these species. To this end, we have devised an automated approach that works to assign precursor charge state from ETD MS/MS data of disulfide-bonded peptides using two Excel-based computational tools.
6.2 REFERENCES


