

# **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

## ABSTRACT

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!!!

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

1.7 Acknowledgements	35
1.8 Summary of subsequent chapters	36
1.9 References	38
<b>2. Collision induced dissociation behavior of <i>N</i>-linked glycopeptides.</b>	<b>50</b>
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 <i>N</i> -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
<b>3. GlycoPep Grader: A web-based utility for assigning the composition of <i>N</i>-linked glycopeptides.</b>	<b>74</b>
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
<b>4. Computational method to determine precursor charge state in ETD MS/MS data of disulfide-bonded peptides.</b>	<b>110</b>
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
<b>5. Future direction: GlycoPep Grader updates.</b>	<b>135</b>
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
<b>6. Conclusion.</b>	<b>159</b>
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:**

Woodin, *et al.* Software for automated interpretation of mass spectrometry data from glycans and glycopeptides. *Analyst*. **2013**, 138, 2793-2803.

### 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.<sup>1, 2, 3, 4, 5</sup> Some of the more frequent PTMs are acetylation, glycosylation, phosphorylation, and disulfide bond formation.<sup>1, 4, 6, 7</sup> Collectively, these modifications work to regulate protein structure and function, and various cellular processes.<sup>1, 2, 7, 8, 9</sup> The development and progression of cancer and other diseases are also shown to be influenced by protein PTMs.<sup>10, 11</sup> 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.<sup>4, 12</sup> Furthermore, both glycosylation and disulfide bond formation have been heavily implicated in the design of safe and effective protein pharmaceuticals.<sup>6, 13, 14</sup>

**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).<sup>15, 16, 17, 18, 19, 20</sup> Currently, mass spectrometry is

the most utilized analytical route in the sequencing of proteins containing PTMs.<sup>16, 21</sup> 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.<sup>22</sup> Furthermore, liquid chromatography mass spectrometry (LC-MS) permits the interrogation of PTMs when heterogeneous protein mixtures are considered.<sup>22, 23, 24</sup> To this end, LC-MS analysis of proteins and peptide PTMs in complex biological matrices is now routinely performed.<sup>22, 23</sup>

**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.<sup>22, 24</sup> This has been accomplished using various types of mass spectrometers.<sup>16, 18, 21, 22, 23, 25</sup> Regardless of the type of mass measurements performed, there are three main components to an MS instrument.<sup>22</sup>

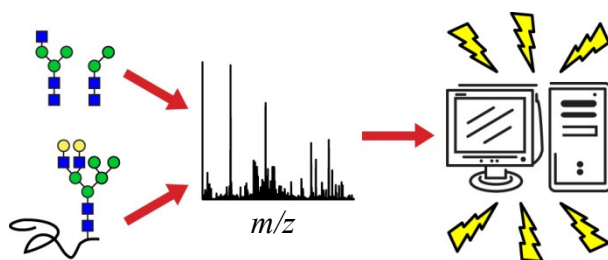
- 1) An ion source to generate ions.
- 2) A mass analyzer to separate ions, based on their  $m/z$ .
- 3) A detector to quantify each ion's abundance.

For the study of biomolecules, electrospray ionization (ESI) and matrix-assisted desorption ionization (MALDI) are the two most common methods used to generate ions.<sup>16, 21, 22,</sup>  
<sup>26</sup> A variety of mass analyzers may be coupled to these ionization sources.<sup>16, 21, 22</sup> ESI is routinely coupled to LC-MS systems with both low resolution, and high resolution, analyzers.<sup>22,</sup>  
<sup>23</sup> 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.<sup>21, 25</sup>

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.<sup>27, 28</sup>

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.<sup>29, 30, 31</sup> 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.<sup>29</sup>

**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.<sup>23, 26, 32, 33, 34, 35, 36</sup> 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.<sup>34, 35, 37, 38</sup>

Droplets containing the charged analyte are formed during the ESI process, which can be performed in positive or negative mode.<sup>16,21</sup> 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.<sup>24</sup>

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.<sup>35,37</sup> 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.<sup>34,35,37</sup> Many theories are proposed for the physical process of droplet formation, including Coulombic fission and solvent evaporation mechanisms.<sup>37</sup> The theories on droplet formation, along with other details of the ESI process, are extensively reviewed in the literature.<sup>35,37</sup>

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.<sup>21</sup> In these instances, other ionization techniques such as MALDI may be more appropriate to use.<sup>39</sup>

**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.<sup>40,41</sup> FT-ICR MS is one type of mass analyzer that provides highly accurate mass information.<sup>21,22,23,40,41,42</sup> For peptides and their associated PTMs, FT-ICR provides powerful MS analysis capabilities.<sup>23,43,44</sup> 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.<sup>41</sup> High resolution FT-ICR MS data is achieved by measuring the cyclotron frequency of ions as they pass through a fixed magnetic field,<sup>40, 41, 42</sup> which is different than the physical contact required for detectors in most other MS instruments.<sup>45</sup>

Although the MS<sup>1</sup> 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.<sup>27, 28</sup>

**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),<sup>22</sup> whereas electron transfer dissociation (ETD) is a relatively new method of tandem mass spectrometry used to study peptides.<sup>46</sup> Both methods result in cleavage along the amino acid backbone of a peptide, though by different mechanisms.<sup>22</sup> In comparison to CID, which cleaves a peptide between the carbonyl and amine groups of adjacent amino acids,<sup>22</sup> ETD cleaves non-specifically along the amide bond of amino acids.<sup>47, 53, 54</sup> ETD allows most labile PTMs to remain intact, while CID produces a signature loss of these modifications.<sup>47</sup> The two dissociation mechanisms provide complementary results to one another, and each offers distinct advantages in the characterization of PTMs.<sup>22, 56</sup> 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.<sup>57, 58</sup> 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.<sup>53, 57, 58, 59</sup> 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.<sup>53,57</sup> 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.<sup>57</sup>

The trajectory properties (kinematics) for these ion collisions have been previously described in detail.<sup>53,59</sup> 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  $KE_{COM}$  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  $KE$  is the kinetic energy of the ion.<sup>57</sup> 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.<sup>53,57</sup>

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

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.<sup>22</sup> Product ions that result from this type of cleavage are termed b- and y-type ions.<sup>22</sup> 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.<sup>22,56</sup>

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.<sup>28</sup>

**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.<sup>46, 50</sup> Unlike the ion/neutral pair activation that occurs in CID, ion/ion pair activation is the basis for ETD fragmentation.<sup>22</sup> ETD Reagents such as fluoranthene provide the radical anions necessary for this transfer of electrons to occur.<sup>46, 50</sup>

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

ETD has also shown to be beneficial in the study of proteins modified by disulfide bonds.<sup>46</sup> For disulfide-bonded peptides, previous work has revealed that ETD preferentially cleaves the disulfide bond between the two joined peptides,<sup>48, 50, 60, 61</sup> leaving a pattern of characteristic product ions that is different in comparison to peptides containing other post-translational modifications.<sup>60</sup> 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.<sup>60</sup> Finally, ETD has been shown to impart more extensive peptide sequence coverage.<sup>47</sup>

## 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.<sup>62, 63, 64, 65</sup> It has long been understood that protein-carbohydrate interactions play a participatory role in many processes affecting disease progression.<sup>62, 65, 66, 67, 68</sup> Furthermore, experimental evidence demonstrates that the identity of the attached glycans change during these events.<sup>67, 69, 70, 71</sup> For example, aberrant glycosylation is often present in individuals experiencing cancer, diabetes, and inflammation.<sup>65, 69, 70, 71, 72</sup> 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.<sup>66, 70, 73, 74</sup>

**1.2.2 Glycosylation Heterogeneity.** Unlike DNA replication and protein transcription, glycosylation is a “non-template”- driven process,<sup>66, 75</sup> where the sugar residues form a multitude of arrangements.<sup>66, 73</sup> 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.<sup>64, 66, 73</sup> This heterogeneity is described in two ways: Glycan differences at different sites of attachment (macroheterogeneity), or within the same site (microheterogeneity).<sup>73</sup> 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.<sup>12</sup> In addition to the established forms of protein glycosylation, including *N*-linked, *O*-linked, and *C*-linked forms,<sup>62, 76, 77, 78</sup> rarer configurations such as *S*-linked

glycosylation,<sup>79</sup> are also being discovered. Although a variety of types exist, the two most common types of glycosylation are *N*-linked and *O*-linked.<sup>62, 68, 80</sup>

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.<sup>73, 78, 81</sup> 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.<sup>81</sup> With *O*-linked glycosylation, the glycan addition may occur at any Ser or Thr residue within the protein sequence,<sup>65, 72</sup> though a very low percentage of these sites are actually occupied.<sup>12</sup> In contrast to *N*-linked glycans, *O*-linked glycans have less defined sequence patterns,<sup>73, 78, 82</sup> and may consist of several distinct core arrangements.<sup>72</sup> For these reasons, both the prediction and determination of *O*-linked glycosylation characteristics have advanced slower than *N*-linked glycosylation analysis.<sup>72</sup>






**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.<sup>80</sup> 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.<sup>80, 83, 84</sup> As it stands, mass spectrometry is considered to be the preferred route of analysis for the identification of protein glycoforms, whether attached or released.<sup>72</sup>

### **1.3 GLYCOSYLATION ANALYSIS BY MASS SPECTROMTERY**

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

through MS experiments.<sup>78, 83, 84</sup> 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.

Monosaccharide	Abbreviation <sup>1</sup>	Mass	Symbol
Fucose	Fuc	146.0579	
Mannose	Hex	162.0528	
Galactose	Hex	162.0528	
<i>N</i> -Acetylglucosamine	HexNAc	203.0794	
Sialic Acid	Neu5Ac	291.0954	

<sup>1</sup> 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.<sup>72, 78, 83</sup> 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,<sup>72, 73</sup> 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.<sup>78, 83, 85</sup> This method is generally advantageous because it provides information about

both glycan composition and the site of the glycan's attachment.<sup>83</sup> 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.<sup>86, 87, 88, 89, 90, 91, 92, 93</sup> For instance, researchers needing to identify occupied *N*-linked glycosylation sites on a specific protein can access UniProtKB,<sup>86</sup> while those wanting statistics specific to proteins modified by *O*-GlcNAc could visit dbOGAP.<sup>93</sup> 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,<sup>87</sup> 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.<sup>85</sup> *N*- and *O*-linked glycans from the same protein can be independently characterized in this manner, as in the method described by Goetz *et al.* where  $\beta$ -elimination is used to release *O*-linked glycans, which are simultaneously permethylated.<sup>94</sup> 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 *et al.*<sup>95</sup> 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.<sup>95</sup> The most likely glycan compositions are selected using precursor mass information.<sup>95</sup> 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.<sup>95</sup> Recently, Goldberg *et al.* extended this concept by developing an automated tool for the analysis of *O*-linked glycans from MS/MS data,<sup>96</sup> 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 *et al.*<sup>97</sup> SysBioWare takes raw MS<sup>1</sup> data that a user uploads and performs baseline adjustment and denoising, wavelet analysis, and peak detection before grouping isotopes of detected peaks.<sup>97</sup> The isotopic grouping is also performed automatically, which enables the program to deduce monoisotopic *m/z* values and precursor

**Table 2.** Glycosylation Databases.

<b>Database</b>	<b>Link to Database</b>	<b>Type</b>	<b>Description</b>
UniProtKB	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>	<i>N</i> -glycos. <i>O</i> -glycos. <i>C</i> -glycos.	Contains annotation of <i>N</i> -, <i>O</i> -, and <i>C</i> -linked glycosylation, as well as glycation. Both mammalian and non-mammalian entries are provided.
dbPTM	<a href="http://dbptm.mbc.nctu.edu.tw/">http://dbptm.mbc.nctu.edu.tw/</a>	<i>N</i> -glycos. <i>O</i> -glycos. <i>C</i> -glycos.	Contains a combinational repertoire of protein PTMs from other databases, including experimentally obtained data on site of modification.
GlycomeDB	<a href="http://www.glycome-db.org/">http://www.glycome-db.org/</a>	<i>N</i> -glycos. <i>O</i> -glycos. <i>C</i> -glycos.	Contains over 30,000 carbohydrate structures from all major taxonomies, representing a variety of glycosylation types.
GlycoSuiteDB	<a href="http://glycosuitedb.expasy.org/glycosuite/glycodb">http://glycosuitedb.expasy.org/glycosuite/glycodb</a>	<i>N</i> -glycos. <i>O</i> -glycos.	Contains over 9400 entries on curated and annotated glycans from a variety of organisms.
O-GlycBase	<a href="http://www.cbs.dtu.dk/databases/OGLYCBASE/">http://www.cbs.dtu.dk/databases/OGLYCBASE/</a>	<i>O</i> -glycos. <i>C</i> -glycos.	Contains over 2000 entries of protein glycosylation sites, the majority of which are <i>O</i> -linked.
UniPep	<a href="http://www.unipep.org/">http://www.unipep.org/</a>	<i>N</i> -glycos.	Contains over 1500 entries of <i>N</i> -linked glycosylation sites found in human proteins.
GlycoBase	<a href="http://glycobase.univ-lille1.fr/base/">http://glycobase.univ-lille1.fr/base/</a>	<i>N</i> -glycos.	Contains HPLC elution positions for 2-AB labeled <i>N</i> -glycans from LC-MS data and exoglycosidase sequencing.
dbOGAP	<a href="http://cbsb.lombardi.georgetown.edu/hulab/OGAP.html">http://cbsb.lombardi.georgetown.edu/hulab/OGAP.html</a>	<i>O</i> -glycos.	Contains over 1100 entries on sites modified by <i>O</i> -GlcNAcylation.

**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.<sup>97</sup> Monosaccharide compositions are then determined by the software on the basis of mass.<sup>97</sup> Currently, the SysBioWare program is being updated to include analysis of MS/MS data for glycans as well.<sup>98</sup> 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.<sup>99</sup> 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.<sup>100</sup> 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 MS<sup>1</sup> data.<sup>101</sup> 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.<sup>101</sup> GlycoSpectrumScan also determines the relative abundance of *N*- and *O*-linked glycoforms that are identified for each glycosylation site.<sup>101</sup> 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.<sup>101</sup> 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.<sup>102</sup> STAT is designed for glycans of up to ten monosaccharide residues, and has the ability to quickly analyze relevant *N*-glycan compositions.<sup>102</sup> 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.<sup>102</sup> Unfortunately, this program is no longer publicly accessible.

An early analysis tool capable of evaluating *O*-linked glycan fragmentation is the OSCAR algorithm.<sup>103</sup> OSCAR, as developed by Ashline *et al.*, is specifically designed for the annotation of permethylated *O*-linked oligosaccharides from MS<sup>n</sup> data.<sup>103</sup> OSCAR is part of a collection of software tools termed Glyspy, which is not currently accessible to the public.<sup>103</sup> Although innovative, the use of OSCAR is limited to direct infusion experiments, as the software does not effectively process data from LC-MS methods.<sup>103</sup>

A program contemporary to OSCAR and also developed to handle glycan MS/MS data is StrOligo.<sup>104</sup> 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.<sup>104</sup> In published research, StrOligo successfully assigned the correct glycan structure in 24 out of 28 cases.<sup>104</sup> 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.<sup>105</sup> The authors describe the web tools GlycoFragment and GlycoSearchMS, which were developed for glycan structural determination.<sup>105</sup> 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.<sup>105</sup> GlycoSearchMS works to analyze experimental glycan data by comparing it against a library of theoretical spectra from *N*-linked and *O*-linked glycan fragmentation entries extracted from SweetDB.<sup>105</sup> 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.<sup>105, 106</sup> 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<sup>1</sup> data, as mentioned previously, but in the identification of glycans from MS/MS data as well.<sup>99</sup> To utilize the glycan fragmentation analysis feature, a user must first input/define the possible glycan compositions and spectral peak list.<sup>99</sup> 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.<sup>99</sup> 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.<sup>107, 108</sup> GlyCH was developed by Tang *et al.* to perform automated interpretation of oligosaccharide tandem mass spectra.<sup>107</sup> The algorithm has a scoring function built in to allow researchers to compare compositions when more than one is determined to be possible.<sup>107</sup> 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.<sup>107</sup> Although not freely accessible online;

the program is available upon request from the authors.<sup>107</sup> More recently, Gao *et al.* developed Glyquest, an automated analysis program that takes a different approach to determine compositions of intact *N*-linked glycans.<sup>108</sup> 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.<sup>108</sup> 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.<sup>108</sup> The glycan compositions with fragmentation profiles that are most similar to the experimental fragmentation are determined to be the most probable candidates.<sup>108</sup> Glyquest is not freely available to the public.

SimGlycan is another program that can be used to increase throughput of glycan analysis.<sup>109, 110</sup> 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.<sup>110</sup> 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.<sup>109</sup> 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.<sup>109</sup> 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.*<sup>96</sup> 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.<sup>96</sup> 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.

MS Analysis Tool	Link to Analysis Tool	Concept and Data Type
GlycoWorkbench	<a href="http://download.glycoworkbench.org/">http://download.glycoworkbench.org/</a>	Identifies and annotates MS and MS/MS data with appropriate glycan compositions or fragments.
GlycanBuilder	<a href="http://live.glycanbuilder.org/">http://live.glycanbuilder.org/</a>	Drawing tool interfacing with GlycoWorkbench that displays different stereochemical representations of glycans.
GlycoSpectrumScan	<a href="http://www.glycospectrumscan.org">http://www.glycospectrumscan.org</a> .	Quantitatively identifies <i>N</i> - and <i>O</i> -linked glycoforms within a protein using LC-MS data.
GlycoFragment	<a href="http://www.glycosciences.de/tools/GlycoFragments/fragment.php4">http://www.glycosciences.de/tools/GlycoFragments/fragment.php4</a>	Identifies and displays the main product ions expected for oligosaccharide MS/MS data.
GlycoSearchMS	<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>	Compares experimental MS/MS data to product ions calculated from an extensive library of <i>N</i> - and <i>O</i> -linked glycans.
SimGlycan	<a href="http://www.premierbiosoft.com/glycan/index.html">http://www.premierbiosoft.com/glycan/index.html</a>	Predicts the structure of glycans from MS/MS data by matching spectra to a built-in database.

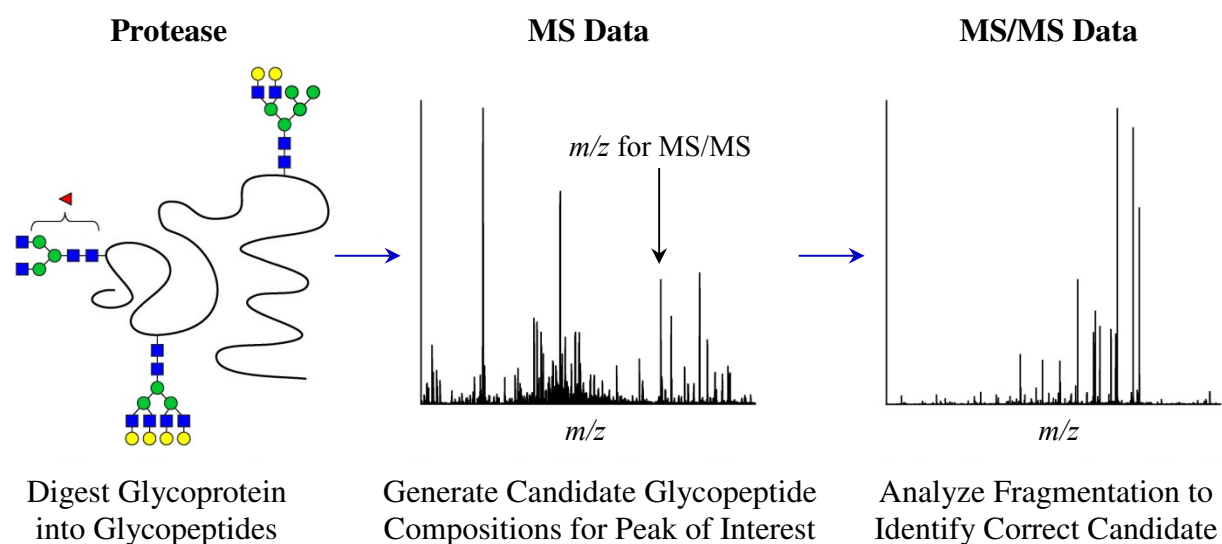
**1.3.2 Automated Analysis of Glycopeptides.** For researchers performing site-specific glycosylation analysis, the initial step toward accomplishing the characterization of attached glycoforms at unique sites within a digested protein is to identify potential glycosylation sites within that protein. 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.<sup>111, 112,</sup>

113, 114, 115, 116, 117, 118, 119 These prediction tools, along with a description and link to each tool, are provided in Table 4.

**Table 4.** Glycosylation Site Prediction Tools.

Database	Link to Database Prediction Tool	Type	Overview
EnsembleGly	<a href="http://turing.cs.iastate.edu/EnsembleGly/">http://turing.cs.iastate.edu/EnsembleGly/</a>	<i>N</i> -glycos. <i>O</i> -glycos. <i>C</i> -glycos.	Uses ensemble learning to predict <i>N</i> -, <i>O</i> -, and <i>C</i> -linked sites, as well as <i>O</i> -glycan types.
GlySeq	<a href="http://www.glycosciences.de/tools/glyseq/">http://www.glycosciences.de/tools/glyseq/</a>	<i>N</i> -glycos. <i>O</i> -glycos.	Uses the PDB and SwissProt to perform statistical analysis of glycosylation sites.
GPP	<a href="http://comp.chem.nottingham.ac.uk/glyco/">http://comp.chem.nottingham.ac.uk/glyco/</a>	<i>N</i> -glycos. <i>O</i> -glycos.	Algorithmically predicts occurrence for <i>N</i> - and <i>O</i> -linked glycosylation.
NetNGlyc	<a href="http://www.cbs.dtu.dk/services/NetNGlyc/">http://www.cbs.dtu.dk/services/NetNGlyc/</a>	<i>N</i> -glycos.	Uses consensus sequence to predict <i>N</i> -linked glycosylation in human proteins.
NetCGlyc	<a href="http://www.cbs.dtu.dk/services/NetCGlyc/">http://www.cbs.dtu.dk/services/NetCGlyc/</a>	<i>C</i> -glycos.	Predicts sites in mammalian proteins for <i>C</i> -mannosylation attachment.
NetOGlyc	<a href="http://www.cbs.dtu.dk/services/NetOGlyc/">http://www.cbs.dtu.dk/services/NetOGlyc/</a>	<i>N</i> -glycos.	Predicts mucin-type GalNAc <i>O</i> -glycosylation in mammalian proteins.
CKSAAP_OGlysite	<a href="http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlysite">http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlysite</a>	<i>O</i> -glycos.	Predicts mucin-type <i>O</i> -glycosylation sites in mammalian proteins.
OGPET	<a href="http://ogpet.utep.edu/">http://ogpet.utep.edu/</a>	<i>O</i> -glycos.	Predicts occurrence of mucin-type <i>O</i> -linked glycosylation in eukaryotic proteins.
YinOYang	<a href="http://www.cbs.dtu.dk/services/YinOYang/">http://www.cbs.dtu.dk/services/YinOYang/</a>	<i>O</i> -glycos.	Predicts <i>O</i> -GlcNAc attachment sites in eukaryotic proteins.
DictyOGlyc	<a href="http://www.cbs.dtu.dk/services/DictyOGlyc/">http://www.cbs.dtu.dk/services/DictyOGlyc/</a>	<i>O</i> -glycos.	Predicts sites for <i>O</i> -GlcNAc attachment in <i>Dictyostelium discoideum</i> proteins.

**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.<sup>27</sup> 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.<sup>29</sup>

## 1.4 AUTOMATED MS and MS<sup>n</sup> 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.<sup>73, 78, 80</sup> This information is useful when deciphering glycopeptide compositions from MS experiments, specifically from CID MS/MS data.<sup>28</sup>

**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.<sup>101, 120, 121</sup> 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.<sup>120</sup> One restriction in the capacity of GlycoMod to analyze glycopeptide data is the inability to handle multiply charged precursors.<sup>120</sup>

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.<sup>121</sup> It also accepts precursor ions in multiple charge states.<sup>121</sup> 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.<sup>121</sup> GlycoSpectrumScan is a more recent program, developed by Deshpande *et al.*, that also interprets MS data on both *N*- and *O*-linked glycopeptides.<sup>101</sup> Like GlycoPep DB, this program has the ability to handle input for both singly and multiply charged data.<sup>101</sup> 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.<sup>27</sup>

**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.<sup>28, 85</sup> These characteristic fragmentation profiles are useful for determining the correct identity of an N-linked glycopeptide when isobaric candidate compositions are possible.<sup>28</sup> 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.<sup>109, 110</sup> 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.<sup>110</sup> 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.<sup>99, 122</sup> On the positive side, Glyco-Peakfinder is useful for *de novo* calculation and annotation of glycan fragment ions within tandem mass spectra.<sup>122</sup> 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).<sup>122</sup>

Additionally, glycan fragmentation is analyzed across multiple charge states, and across multiple charge carriers (cationic carriers), within the same spectrum.<sup>122</sup>

A completely different approach is used in GlycoPep ID.<sup>123</sup> 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.<sup>123</sup> 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.<sup>123</sup>

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 Peptonist, Medice Integrator, the Branch-and-Bound algorithm, GlycoMaster, Sweet Substitute, and GlyDB.<sup>124, 125, 126, 127, 128, 129</sup> 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.*<sup>130</sup> In the

analysis of 3132 spectra, the software was reported to have found 338 that corresponded to MS/MS data of glycopeptides (versus peptides).<sup>130</sup> 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.<sup>130</sup> However, GlycoMiner is only capable of performing compositional analysis when the spectra are of good quality.<sup>130</sup> The program fails when spectral quality is low, as evidenced by the software's identification of glycan composition in only 196/338 glycopeptide spectra.<sup>130</sup> 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.<sup>130</sup> 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.<sup>130</sup> 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.<sup>131</sup> 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.<sup>131</sup> To generate the peptide-glycan pairs, GPS must find both low mass oxonium, and *N*-glycan core-containing, product ions.<sup>131</sup> GPS is freely available online, as well.<sup>131</sup> 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<sup>1</sup> and MS<sup>2</sup> information extracted from LC-MS/MS experiments.<sup>132</sup> One of the benefits to the

method is that no prior knowledge of the potential glycosylation or identity of the glycopeptide is necessary.<sup>132</sup> Instead, GlypID assigns a cluster of glycopeptides in the “same family” (microheterogeneities) based on observed mass.<sup>132</sup> In addition, the approach utilizes an isotope deconvolution algorithm to assign ion charges along with monoisotopic ions.<sup>132</sup> This information is then added to the inclusion list of “prioritized precursor ions” for the MS/MS analysis that follows.<sup>132</sup> Next, the resultant CID data is searched for the longest series of glycosidic bond cleavage series.<sup>132</sup> 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.<sup>132</sup> A score is assigned to the CID spectrum based on this sequence tag.<sup>132</sup> 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.<sup>132</sup> 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.<sup>132</sup> 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.<sup>133</sup> 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.<sup>134</sup> The new software tool, GlypID 2.0, uses high resolution MS<sup>1</sup> data along with CID and HCD scan information to improve the accuracy of *N*-linked glycopeptide

identification.<sup>134</sup> Like the original GlypID, GlypID 2.0 can also score CID spectra independently on MS systems that do not contain the HCD instrument option.<sup>134</sup> 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.<sup>28</sup> 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.<sup>28</sup>

GPG utilizes the MS/MS data by calculating, scoring, and searching for the expected product ions of potential glycopeptide candidate compositions.<sup>28</sup> 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.<sup>28</sup> 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.<sup>28</sup>

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.<sup>28</sup> 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.<sup>28</sup> 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 MS<sup>2</sup> data with GPG.<sup>28</sup> 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.

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

**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.<sup>68, 80,</sup>

<sup>135</sup> 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.<sup>82</sup>

**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 *N*-acetylgalactosamine (GalNAc) residues to the hydroxyl group of Ser/Thr side chains (commonly referred to as the Tn antigen).<sup>63,68</sup> 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.<sup>101</sup> GlycoSpectrumScan is designed to analyze LC-MS data of intact glycopeptides from proteolytic digests.<sup>101</sup> The program utilizes MS<sup>1</sup> data to determine glycopeptide composition, along with the relative distribution of glycoforms at each of the sites.<sup>101</sup> 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](http://www.glycospectrumscan.org)).<sup>101</sup>

GlycoX and GlycoMod, described earlier in the analysis of *N*-linked glycopeptides, are capable of *O*-linked glycopeptide data interpretation as well.<sup>120,136</sup> Unlike GlycoMod (<http://web.expasy.org/glycomod/>), GlycoX is not publicly available, though it is available upon request from the authors.<sup>136</sup> GlycoWorkbench, also described previously, performs automation

of *O*-linked glycopeptide MS<sup>1</sup> data to elucidate the most likely composition from an experimental peak list in the same manner as for *N*-linked glycopeptide spectra.<sup>99</sup>

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.<sup>99</sup> 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.<sup>99</sup>

There are promising advances being made in the compositional determination of glycopeptides using ETD fragmentation techniques,<sup>72, 84</sup> or a combination of CID and ETD, particularly in the study of *O*-linked species.<sup>137</sup> A recent method described by Darula *et al.* in which MS<sup>1</sup>, CID, and ETD data are used in conjunction with Protein Prospector v5.3 for the identification of SA<sub>1-10</sub>GalGalNAc-containing *O*-linked glycopeptides enriched from bovine serum, demonstrates the potential for automated analysis through a combination of these techniques and database searches.<sup>137</sup> However, this process is only semi-automated, and restricted to samples containing simple carbohydrate structures.<sup>137</sup> 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

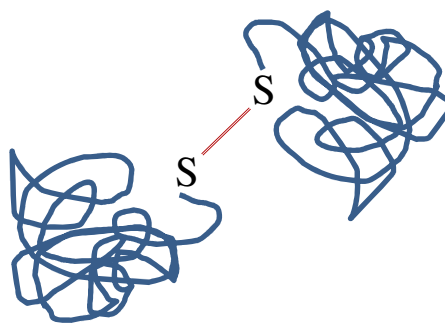
**1.5.1 Overview of Disulfide Bonds.** Protein disulfide bonds result when two cysteine residues are covalently joined through oxidation.<sup>1, 4, 7</sup> The formation of native disulfide bonds in a protein is necessary to achieve proper folding, stability and conformation.<sup>1, 8, 9, 14, 15</sup>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

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.<sup>6, 13, 15, 138, 139</sup> For many of these, *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 *E. coli*.<sup>139</sup>

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 *et al.*, the demand for this characterization is rapidly increasing,<sup>14</sup> 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.<sup>18</sup> 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.** Intergain 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.<sup>18, 50, 140</sup>

**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.<sup>141, 142</sup> 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.<sup>18</sup>

**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.<sup>18, 23</sup> 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.<sup>55</sup> Of these MS systems, not all provide high resolution MS<sup>1</sup> scans.<sup>23, 55</sup> With low resolution mass spectrometers, it is often necessary to assign charge state independent of the MS<sup>1</sup> 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.<sup>56, 142</sup> 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.<sup>22</sup> 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,<sup>22</sup> 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.<sup>143, 144, 145, 146, 147, 148</sup> 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.<sup>149, 150, 151</sup>

## **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

- (1) Kang, T. S.; Kini, R. M. Structural determinants of protein folding. *Cell. Mol. Life Sci.* **2009**, *66*, 2341-2361.
- (2) Jensen, O. N. Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 391-403.
- (3) Sato, Y.; Inaba, K. Disulfide bond formation network in the three biological kingdoms, bacteria, fungi and mammals. *Febs J.* **2012**, *279*, 2262-2271.
- (4) Wong, J. W. H.; Ho, S. Y. W.; Hogg, P. J. Disulfide bond acquisition through eukaryotic protein evolution. *Mol. Biol. Evol.* **2011**, *28*, 327-334.
- (5) Reinders, J. Sickmann, A. Modificomics: Posttranslational modifications beyond protein phosphorylation and glycosylation. *Biomolecular Engineering.* **2007**, *24*, 169-177.
- (6) Walsh, G.; Jefferis, R. Post-translational modifications in the context of therapeutic proteins. *Nat. Biotechnol.* **2006**, *24*, 1241-1252.
- (7) Walsh, C. T.; Garneau-Tsodikova, S.; Gatto, G. J. Protein posttranslational modifications: The chemistry of proteome diversifications. *Angew. Chem. Int. Edit.* **2005**, *44*, 7342-7372.
- (8) Braakman, I.; Bulleid, N. J. Protein folding and modification in the mammalian endoplasmic reticulum. *Annu. Rev. Biochem.* **2011**, *80*, 71-99.
- (9) Fass, D. Disulfide bonding in protein biophysics. *Annu. Rev. Biophys.* **2012**, *41*, 63-79.
- (10) Krueger, K. E.; Srivastava, S. Posttranslational protein modifications – Current implications for cancer detection, prevention, and therapeutics. *Mol. Cell. Proteomics.* **2006**, *5*, 1799-1810.
- (11) Karve, T. M.; Cheema, A. K. Small changes huge impact: The role of protein posttranslational modifications in cellular homeostasis and disease. *J. Amino Acids.* **2011**, *2011*.
- (12) Apweiler, R.; Hermjakob, H.; Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *J. Biochim. Biophys. Acta.* **1999**, *1473*, 4-8.
- (13) Jenkins, N.; Murphy, L.; Tyther, R. Post-translational modifications of recombinant proteins: Significance for biopharmaceuticals. *Mol. Biotechnol.* **2008**, *39*, 113-118.
- (14) Trivedi, M. V.; Laurence, J. S.; Siahaan, T. J. The role of thiols and disulfides on protein stability. *Curr. Protein Pept. Sci.* **2009**, *10*, 614-625.
- (15) Woycechowsky, K. J.; Raines, R. T. Native disulfide bond formation in proteins. *Curr. Opin. Chem. Biol.* **2000**, *4*, 533-539.



- (16) Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21*, 255-261.
- (17) Hoffman, M. D.; Sniatynski, M. J.; Kast, J. Current approaches for global post-translational modification discovery and mass spectrometric analysis. *Anal. Chim. Acta.* **2008**, *627*, 50-61.
- (18) Gorman, J. J.; Wallis, T. P.; Pitt, J. J. Protein disulfide bond determination by mass spectrometry. *Mass Spectrom. Rev.* **2002**, *21*, 183-216.
- (19) Mobli, M.; King, G. F. NMR methods for determining disulfide-bond connectivities. *Toxicon.* **2010**, *56*, 849-854.
- (20) McGeehan, J. E.; Bourgeois, D.; Royant, A.; Carpentier, P. Raman-assisted crystallography of biomolecules at the synchrotron: Instrumentation, methods and applications. *BBA-Proteins Proteomics.* **2011**, *1814*, 750-759.
- (21) Mann, M.; Hendrickson, R. C.; Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **2001**, *70*, 437-473.
- (22) Han, X.; Aslanian, A.; Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **2008**, *12*, 483-490.
- (23) Meng, F.; Forbes, A. J.; Miller, L. M.; Kelleher, N. L. Detection and localization of protein modifications by high resolution tandem mass spectrometry. *Mass Spectrom. Rev.* **2005**, *24*, 126-134.
- (24) Witze, E. S.; Old, W. M.; Resing, K. A.; Ahn, N. G. Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods.* **2007**, *4*, 798-806.
- (25) Chaurand, P.; Luetzenkirchen, F.; Spengler, B. Peptide and protein identification by matrix-assisted laser desorption ionization (MALDI) and MALDI-post-source decay time-of-flight mass spectrometry. *Mass Spectrom. Rev.* **1999**, *10*, 91-103.
- (26) Trauger, S. A.; Webb, W.; Siuzdak, G. Peptide and protein analysis with mass spectrometry. *Spectroscopy.* **2002**, *16*, 15-28.
- (27) Desaire, H.; Hua, D. When can glycopeptides be assigned based solely on high-resolution mass spectrometry data? *Int. J. Mass. Spectrom.* **2009**, *287*, 21-26.
- (28) Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. GlycoPep Grader: A web-based utility for assigning the composition of N-linked glycopeptides. *Anal. Chem.* **2012**, *84*, 4821-4829.
- (29) Woodin, C. L.; Maxon, M.; Desaire, H. Software for automated interpretation of mass spectrometry data from glycans and glycopeptides. *Analyst.* **2013**, *138*, 2793-2803.

- (30) Ahrné, E.; Müller, M.; Lisacek, F. Unrestricted identification of modified proteins using MS/MS. *Proteomics*. **2010**, *10*, 671-686.
- (31) Matthiesen, R.; Trelle, M. B.; Højrup, P.; Bunkenborg, J.; Jensen, O. N. VEMS 3.0: Algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Proteome Res.* **2005**, *4*, 2338-2347.
- (32) Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci.* **1986**, *83*, 6233-6237.
- (33) Biemann, K.; Papayannopoulos, I. A. Amino acid sequencing of proteins. *Acc. Chem. Res.* **1994**, *27*, 370-378.
- (34) Griffiths, W. J.; Jonsson, A. P.; Liu, S.; Rai, D. K.; Wang, Y. Electrospray and tandem mass spectrometry in biochemistry. *Biochem. J.* **2001**, *355*, 545-561.
- (35) Smith, R. D.; Loo, J. A.; Edmonds, C. G.; Barinaga, C. J.; Udseth, H. R. New developments in biochemical mass spectrometry: Electrospray ionization. *Anal. Chem.* **1990**, *62*, 882-899.
- (36) Weiskopf, A. S.; Vouros, P.; Harvey, D. J. Electrospray ionization-ion trap mass spectrometry for structural analysis of complex *N*-linked glycoprotein oligosaccharides. *Anal. Chem.* **1998**, *70*, 4441-4447.
- (37) Gabelica, V. De Pauw, E. Internal energy and fragmentation of ions produced in electrospray sources. *Mass Spectrom. Rev.* **2005**, *24*, 566-587.
- (38) Chassigne, H.; Vacchina, V.; Łobinski, R. Elemental speciation analysis in biochemistry by electrospray mass spectrometry. *Trends in Analytical Chemistry*. **2000**, *19*, 300-313.
- (39) Hardouin, J. Protein sequence information by matrix-assisted laser desorption/ionization in-source decay spectrometry. *Mass Spectrom. Rev.* **2007**, *26*, 672-682.
- (40) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.* **1998**, *17*, 1-35.
- (41) Schrader, W.; Klein, H. W. Liquid chromatography/Fourier transform ion cyclotron resonance mass spectrometry (LC-FTICR MS): An early overview. *Anal. Bioanal. Chem.* **2004**, *379*, 1013-1024.
- (42) Laskin, J.; Futrell, J. H. Collisional activation of peptide ions in FT-ICR mass spectrometry. *Mass Spectrom. Rev.* **2003**, *22*, 158-181.
- (43) Emmett, M. R. Determination of post-translational modifications of proteins by high-sensitivity, high-resolution Fourier transform ion cyclotron resonance mass spectrometry. *J. Chromatogr. A.* **2003**, *1013*, 203-213.

- (44) Morelle, W.; Michalski, J. C. The mass spectrometric analysis of glycoproteins and their glycan structures. *Curr. Anal. Chem.* **2005**, *1*, 29-57.
- (45) Glish, G. L.; Vachet, R. W. The basics of mass spectrometry in the twenty-first century. *Nat. Rev. Drug Disc.* **2003**, *2*, 140-150.
- (46) Mikesch, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E. P.; Shabanowitz, J.; Hunt, D. F. The utility of ETD mass spectrometry in proteomic analysis. *BBA-Proteins Proteomics.* **2006**, *1764*, 1811-1822.
- (47) Molina, H.; Matthiessen, R.; Kandasamy, K.; Pandey, A. Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Anal. Chem.* **2008**, *80*, 4825-4835.
- (48) Wiesner, J.; Premisler, T.; Sickmann, A. Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics.* **2008**, *8*, 4466-4483.
- (49) Jones, A. W.; Cooper, H. J. Dissociation techniques in mass spectrometry-based proteomics. *Analyst.* **2011**, *136*, 3419-3429.
- (50) Cole, S. R.; Ma, X.; Zhang, X.; Xia, Y. Electron transfer dissociation (ETD) of peptides containing intrachain disulfide bonds. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 310-320.
- (51) Guthals, A.; Bandeira, N. Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Mol. Cell. Biol.* **2012**, *11*, 550-557.
- (52) Sun, R. X.; Dong, M. Q.; Song, C. Q.; Chi, H.; Yang, B.; Xiu, L. Y.; Tao, L.; Jing, Z. Y.; Liu, C.; Wang, L. H.; Fu, Y.; He, S. M. Improved peptide identification for proteomic analysis based on comprehensive characterization of electron transfer dissociation spectra. *J. Proteome Res.* **2010**, *9*, 6354-6367.
- (53) Sleno, L.; Volmer, D. A. Ion activation methods for tandem mass spectrometry. *J. Mass Spectrom.* **2004**, *39*, 1091-1112.
- (54) Sobott, F. Watt, S. J.; Smith, J.; Edelman, M. J.; Kramer, H. B.; Kessler, B. M. Comparison of CID versus ETD based MS/MS fragmentation for the analysis of protein ubiquitination. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1652-1659.
- (55) Xia, Q.; Lee, M. V.; Rose, C. M.; Marsh, A. J.; Hubler, S. L.; Wenger, C. D.; Coon, J. J. Characterization and diagnostic value of amino acid side chain neutral losses following electron-transfer dissociation. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 255-264.
- (56) Larsen, M. R.; Trelle, M. B.; Thingholm, T. E.; Jensen, O. N. Analysis of posttranslational modifications of proteins by tandem mass spectrometry. *Biotechniques.* **2006**, *40*, 790-798.
- (57) Wells, J. M.; McLuckey, S. A. Collision-induced dissociation (CID) of peptides and

proteins. *Methods Enzymol.* **2005**, *402*, 148-185.

(58) Smith, R. D.; Loo, J. A.; Barinaga, C. J.; Edmonds, C. G.; Udseth, H. R. Collisional activation and collision-activated dissociation of large multiply charged polypeptides and proteins produced by electrospray ionization. *J. Am. Soc. Mass Spectrom.* **1990**, *1*, 53-65.

(59) Shukla, A. K.; Futrell, J. H. Tandem mass spectrometry: Dissociation of ions by collisional activation. *J. Am. Soc. Mass Spectrom.* **2000**, *1*, 6-15.

(60) Wu, S. L.; Jiang, H.; Lu, Q.; Dai, S.; Hancock, W. S.; Karger, B. L. Mass spectrometric determination of disulfide linkages in recombinant therapeutic proteins using online LC-MS with electron-transfer dissociation. *Anal. Chem.* **2009**, *81*, 112-122.

(61) Wang, Y.; Lu, Q.; Wu, S. L.; Karger, B. L.; Hancock, W. S. Characterization and comparison of disulfide linkages and scrambling patterns in therapeutic monoclonal antibodies: Using LC-MS with electron transfer dissociation. *Anal. Chem.* **2011**, *83*, 3133-3140.

(62) Spiro, R. G. Protein glycosylation: Nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology.* **2002**, *12*, 43R-56R.

(63) Hang, H.; Bertozzi, C. R. The chemistry and biology of mucin-type *O*-linked glycosylation. *Bioorg. Med. Chem.* **2005**, *13*, 5021-5034.

(64) Murrell, M. P.; Yarema, K. J.; Levchenko, A. The systems biology of glycosylation. *Chem. Biochem.* **2004**, *5*, 1334-1347.

(65) Van den Steen, P.; Rudd, P. M.; Dwek, R. A.; Opdenakker, G. Concepts and principles of *O*-linked glycosylation. *Crit. Rev. Biochem. Mol. Biol.* **1998**, *33*, 151-208.

(66) Bertozzi, C. R.; Kiessling, L. L. Chemical glycobiology. *Science.* **2001**, *291*, 2357-2364.

(67) Dennis, J. W.; Granovsky, M.; Warren, C. Protein glycosylation in development and disease. *BioEssays.* **1999**, *21*, 412-421.

(68) Tian, E. and Ten Hagen, K. G. Recent insights into the biological roles of mucin-type *O*-glycosylation. *Glycoconjugate J.* **2009**, *26*, 325-334.

(69) Wada, Y.; Tajiri, M.; Ohshima, S. Quantitation of saccharide compositions of *O*-glycans by mass spectrometry of glycopeptides and its application to rheumatoid arthritis. *J. Proteome Res.* **2010**, *9*, 1367-1373.

(70) Dube, D. H.; Bertozzi, C. R. Glycans in cancer and inflammation – Potential for therapeutics and diagnostics. *Nat. Rev. Drug Discovery.* **2005**, *4*, 477-488.

(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*-

GlcNAcylation is involved in the etiology of cardiovascular disorders, type-2 diabetes and Alzheimer's disease. *J. Biochim. Biophys. Acta.* **2010**, *1800*, 67-79.

(72) Jensen, P. H.; Kolarich, D.; Packer, N. H. Mucin-type *O*-glycosylation – Putting the pieces together. *Febs J.* **2010**, *277*, 81-94.

(73) Mariño, K.; Bones, J.; Kattla, J. J.; Rudd, P. M. A systematic approach to protein glycosylation analysis: A path through the maze. *Nat. Chem. Biol.* **2010**, *6*, 713-723.

(74) Budnik, B. A.; Lee, R. S.; Steen, J. A. J. Global methods for protein glycosylation analysis by mass spectrometry. *J. Biochim. Biophys. Acta.* **2006**, *1764*, 1870-1880.

(75) Raman, R.; Raguram, S.; Venkataraman, G.; Paulson, J. C.; Sasiekharan, R. Glycomics: An integrated systems approach to structure-function relationships of glycans. *Nat. Methods.* **2005**, *2*, 817-824.

(76) Brazier-Hicks, M.; Evans, K. M.; Gershater, M. C.; Puschmann, H.; Steel, P. G.; Edwards, R. The *C*-glycosylation of flavonoids in cereals. *J. Biol. Chem.* **2009**, *284*, 17926-17934.

(77) Hofsteenge, J.; Müller, D. R.; de Beer, T.; Löffler, A.; Richter, W. J.; Vliegthart, J. F. G. New type of linkage between a carbohydrate and a protein: *C*-glycosylation of a specific tryptophan residue in human RNase U<sub>s</sub>. *Biochemistry.* **1994**, *33*, 13524-13530.

(78) Morelle, W.; Canis, K.; Chirat, F.; Faid, V.; Michalski, J. C. The use of mass spectrometry for the proteomic analysis of glycosylation. *Proteomics.* **2006**, *6*, 3993-4015.

(79) Stepper, J.; Shasti, S.; Loo, T. S.; Preston, J. C.; Novak, P.; Man, P.; Moore, C. H.; Havlí ek, V.; Patchett, M. L.; Norris, G. E. Cysteine *S*-glycosylation, a new post-translational modification found in glycopeptide bacteriocins. *FEBS Lett.* **2011**, *585*, 645-650.

(80) Rakus, J. F.; Mahal, L. K. New technologies for glycomic analysis: Toward a systematic understanding of the glycome. *Ann. Rev. Anal. Chem.* **2001**, *4*, 367-392.

(81) Jones, J.; Krag, S. S.; Betenbaugh, M. J. Controlling *N*-glycan site occupancy. *Biochim. Biophys. Acta.* **2005**, *1726*, 121-137.

(82) Christlet, T. H. T.; Veluraja, K. Database analysis of *O*-glycosylation sites in proteins. *Biophys. J.* **2001**, *80*, 952-960.

(83) Dalpathado, D. S.; Desaire, H. Glycopeptide analysis by mass spectrometry. *Analyst.* **2008**, *133*, 731-738.

(84) North, S. J.; Hitchen, P. G.; Haslam, S. M.; Dell, A. Mass spectrometry in the analysis of *N*-linked and *O*-linked glycans. *Curr. Opin. Struct. Biol.* **2009**, *19*, 498-506.

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

B. J.; Lebrilla, C. B. Simultaneous and extensive site-specific *N*- and *O*-glycosylation analysis in protein mixtures. *J. Proteome Res.* **2011**, *10*, 2612-2624.

(86) Wu, C. H.; Apweiler, R.; Bairoch, A.; Natale, D. A.; Barker, W. C.; Boeckermann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Mazumder, R.; O'Donovan, C.; Redaschi, N.; Suzek, B. The Universal Protein Resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res.* **2006**, *34*, D187-D191.

(87) Lee, T. Y.; Huang, H. D.; Hung, J. H.; Huang, H. Y.; Yang, Y. S.; Wang, T. H. dbPTM: An informal repository of protein post-translational modification. *Nucleic Acids Res.* **2006**, *34*, D622-D627.

(88) Ranzinger, R.; Herget, S.; Wetter, T.; von der Lieth, C. W. Glycome DB – Integration of open-access carbohydrate structure search databases. *BMC Bioinformatics.* **2008**, *9*.

(89) Cooper, C. A.; Harrison, M. J.; Wilkins, M. R.; Packer, N. H. GlycoSuiteDB: A new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.* **2001**, *29*, 332-335.

(90) Gupta, R.; Birch, H.; Rapacki, K.; Brunak, S.; Hansen, J. E. O-GLYCBASE version 4.0: A revised database of *O*-glycosylated proteins. *Nucleic Acids Res.* **1999**, *27*, 370-372.

(91) Zhang, H.; Loriaux, P.; Eng, J.; Campbell, D.; Keller, A.; Moss, P.; Bonneau, R.; Zhang, N.; Zhou, Y.; Wollscheid, B.; Cooke, K.; Yi, E. C.; Lee, H.; Peskind, E. R.; Zhang, J.; Smith, R. D.; Aebersold, R. UniPep – A database for human *N*-linked glycosites: A resource for biomarker discovery. *Genome Biol.* **2006**, *7*.

(92) Campbell, M. P.; Roylez, L.; Radcliffe, C. M.; Dwek, R. A.; Rudd, P. M. GlycoBase and autoGU: Tools for HPLC-based glycan analysis. *Bioinformatics.* **2008**, *24*, 1214-1216.

(93) Wang, J.; Torii, M.; Liu, H.; Hart, G. W.; Hu, Z. Z. dbOGAP – An integrated bioinformatics resource for protein *O*-GlcNAcylation. *BMC Bioinformatics.* **2011**, *12*.

(94) Goetz, J. A.; Novotny, M. V.; Mechref, Y. Enzymatic/chemical release of *O*-glycans allowing MS analysis at high sensitivity. *Anal. Chem.* **2009**, *81*, 9546-9552.

(95) Goldberg, D.; Sutton-Smith, M.; Paulson, J.; Dell, A. Automatic annotation of matrix-assisted laser desorption/ionization *N*-glycan spectra. *Proteomics.* **2005**, *5*, 865-876.

(96) Goldberg, D.; Bern, M.; Li, B.; Lebrilla, C. B. Automatic determination of *O*-glycan structure from fragmentation spectra. *J. Proteome Res.* **2006**, *5*, 1429-1434.

(97) Vakhrushev, S. Y.; Dadimov, D.; Peter-Katalini, J. Software platform for high-throughput glycomics. *Anal. Chem.* **2009**, *81*, 3252-3260.

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

for automated glycomics. *Glyco-Bioinformatics*. Postdam, Germany; **2009**, 141-161.

(99) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. GlycoWorkBench: A tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.* **2008**, *7*, 1650-1659.

(100) Damerell, D.; Ceroni, A.; Maass, K.; Ranzinger, R.; Dell, A.; Haslam, S. M. The GlycanBuilder and GlycoWorkbench glycoinformatics tools: Updates and new developments. *Biol. Chem.* **2012**, *393*, 1357-1362.

(101) Deshpande, N.; Jensen, P. H.; Packer, N. H.; Kolarich, D. GlycoSpectrumScan: Fishing glycopeptides from MS spectra of protease digests of human colostrum sIgA. *J. Proteome Res.* **2010**, *9*, 1063-1075.

(102) Gaucher, S. P.; Morrow, J.; Leary, J. STAT: A saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal. Chem.* **2000**, *72*, 2331-2336.

(103) Ashline, D. J.; Lapadula, A. J.; Liu, Y. H.; Lin, M.; Grace, M.; Pramanik, B.; Reinhold, V. N. Carbohydrate structural isomers analyzed by sequential mass spectrometry. *Anal. Chem.* **2007**, *79*, 3830-3842.

(104) Ethier, M.; Saba, J. A.; Ens, W.; Standing, K. G.; Perreault, H. Automated structural assignment of derivatized complex *N*-linked oligosaccharides from tandem mass spectra. *Rapid Commun. Mass. Spectrom.* **2002**, *16*, 1743-1754.

(105) Lohmann, K. K.; von der Lieth, C. W. GlycoFragment and GlycoSearchMS: Web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res.* **2004**, *32*, W261-W266.

(106) Lohmann, K. K.; von der Lieth, C. W. GLYCO-FRAGMENT: A web tool to support the interpretation of mass spectra of complex carbohydrates. *Proteomics.* **2003**, *3*, 2028-2035.

(107) Tang, H.; Mechref, Y.; Novotny, M. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics.* **2005**, *21*, I431-I439.

(108) Gao, H. Y. Generation of asparagine-linked glycan structure databases. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1739-1742.

(109) Blow, N. A spoonful of sugar. *Nature.* **2009**, *457*, 617-620.

(110) Apte, A.; Meitei, N. S. Bioinformatics in glycomics: Glycan characterization with mass spectrometric data using SimGlycan<sup>TM</sup>. *Methods Mol. Biol.* **2009**, *600*, 269-281.

(111) Caragea, C.; Sinapov, J.; Silvescu, A.; Dobbs, D.; Honavar, V. Glycosylation site prediction using ensembles of support vector machine classifiers. *BMC Bioinformatics.* **2007**, *8*.

- (112) Lütteke, T.; Frank, M.; von der Lieth, C. W. Carbohydrate Structure Suite (CSS): Analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res.* **2005**, *33*, D242-D246.
- (113) Hamby, S. E.; Hirst, J. D. Prediction of glycosylation sites using random forests. *BMC Bioinformatics.* **2008**, *9*.
- (114) Gupta, R.; Brunak, S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput.* Lyngby, Denmark; **2002**, 310-322.
- (115) Julenius, K. NetCGlyc 1.0: Prediction of mammalian C-mannosylation sites. *Glycobiology.* **2007**, *17*, 868-876.
- (116) Hansen, J. E.; Lund, O.; Tolstrup, N.; Gooley, A. A.; Williams, K. L.; Brunak, S. NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate J.* **1998**, *15*, 115-130.
- (117) Chen, Y. Z.; Tang, Y. R.; Sheng, Z. Y.; Zhang, Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of *k*-spaced amino acid pairs. *BMC Bioinformatics.* **2008**, *9*.
- (118) Gerken, T. A.; Jamison, O.; Perrine, C. L.; Collette, J. C.; Moinova, H.; Ravi, L.; Markowitz, S. D.; Shen, W.; Patel, H.; Tabak, L. A. Emerging paradigms for the initiation of mucin-type protein O-glycosylation by the polypeptide GalNAc transferase family of glycosyltransferases. *J. Biol. Chem.* **2011**, *286*, 14493-14507.
- (119) Gupta, R.; Jung, E.; Gooley, A. A.; Williams, K. L.; Brunak, S.; Hansen, J. Scanning the available *Dictyostelium discoideum* proteome for O-linked GlcNAc glycosylation sites using neural networks. *Glycobiology.* **1999**, *9*, 1009-1022.
- (120) Cooper, C. A.; Gasteiger, E.; Packer, N. H. GlycoMod – A software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics.* **2001**, *1*, 340-349.
- (121) Go, E. P.; Rebecchi, K. R.; Dalpathado, D. S.; Bandu, M. L.; Zhang, Y.; Desaire, H. GlycoPep DB: A tool for glycopeptide analysis using a “smart search”. *Anal. Chem.* **2007**, *79*, 1708-1713.
- (122) Maass, K.; Ranzinger, R.; Geyer, H.; von der Lieth, C. W.; Geyer, R. “Glyco-peakfinder” – *De novo* composition analysis of glycoconjugates. *Proteomics.* **2007**, *7*, 4435-4444.
- (123) Irungu, J.; Go, E. P.; Dalpathado, D. S.; Desaire, H. Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID. *Anal. Chem.* **2007**, *79*, 3065-3074.
- (124) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H. R.; Dell, A. Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J. Proteome Res.* **2007**, *6*, 3995-4005.



- (125) Joenväärä, S.; Ritamo, I.; Peltoniemi, H.; Renkonen, R. *N*-glycoproteomics – An automated workflow approach. *Glycobiology*. **2008**, *18*, 339-349.
- (126) Peltoniemi, H.; Joenväärä, S.; Renkonen, R. *De novo* glycan structure search with the CID MS/MS spectra of native *N*-glycopeptides. *Glycobiology*. **2009**, *19*, 707-714.
- (127) Shan, B.; Ma, B.; Zhang, K. Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *J. Bioinform. Comput. Biol.* **2008**, *6*, 77-91.
- (128) Clerens, S.; den Ende, V. W.; Verhaet, P.; Geenen, L.; Archens, L. Sweet Substitute: A software tool for *in silico* fragmentation of peptide-linked *N*-glycans. *Proteomics*. **2004**, *4*, 629-632.
- (129) Ren, J. M.; Rejtar, T.; Li, L.; Karger, B. L. *N*-glycan structure annotation of glycopeptides using a linearized glycan structure database (GlyDB). *J. Proteome Res.* **2007**, *6*, 3162-3173.
- (130) Ozohanics, O.; Krenyacz, J.; Ludányi, K.; Pollreisz, F.; Vékey, K.; Drahos, L. GlycoMiner: A new software tool to elucidate glycopeptide composition. *Rapid Commun. Mass. Spectrom.* **2008**, *22*, 3245-3254.
- (131) Pompach, P.; Chandler, K.; Lan, R.; Edwards, N.; Goldman, R. Semi-automated identification of *N*-glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search. *J. Proteome Res.* **2012**, *11*, 1728-1740.
- (132) Wu, Y.; Mechref, Y.; Klouckova, I.; Mayampurath, A. M.; Novotny, M. V.; Tang, H. Mapping site-specific protein *N*-glycosylations through liquid chromatography/mass spectrometry and targeted tandem mass spectrometry. *Rapid Commun. Mass. Spectrom.* **2010**, *24*, 965-972.
- (133) Wu, Y.; Mechref, Y.; Klouckova, I.; Novotny, M. V.; Tang, H. A computational approach for the identification of site-specific protein glycosylations through ion-trap mass spectrometry. *Syst. Biol. Comput. Prot.* Berlin/Heidelberg, Germany: **2007**; *4532*, 96-107.
- (134) Mayampurath, A. M.; Wu, Y.; Segu, Z. M.; Mechref, Y.; Tang, H. Improving confidence in detection and characterization of protein *N*-glycosylation sites and microheterogeneity. *Rapid Commun. Mass. Spectrom.* **2011**, *25*, 2007-2019.
- (135) Mazola, Y.; China, G.; Mussacchio, A. Integrating bioinformatics tools to handle glycosylation. *PLOS Comput. Biol.* **2011**, *7*.
- (136) An, H. J.; Tillinghast, J. S.; Woodruff, D. L.; Rocke, D. M.; Lebrilla, C. B. A new computer program (GlycoX) to determine simultaneously the glycosylation sites and oligosaccharide heterogeneity of glycoproteins. *J. Proteome Res.* **2006**, *5*, 2800-2808.
- (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

glycopeptides. *Eur. J. Mass Spectrom.* **2010**, *16*, 421-428.

(138) Kálmán-Szekeres, Z.; Olajos, M.; Ganzler, K. Analytical aspects of biosimilarity issues of protein drugs. *J. Pharm. Biomed. Anal.* **2012**, *69*, 185-195.

(139) Barnes, C. A. S.; Lim, A. Applications of mass spectrometry for the structural characterization of recombinant pharmaceuticals. *Mass Spectrom. Rev.* **2007**, *26*, 370-388.

(140) Koivu, J.; Myllylä, R. Interchain disulfide bond formation in types I and II procollagen. *J. Biol. Chem.* **1987**, *13*, 6159-6164.

(141) Yen, T. Y.; Joshi, R. K.; Yan, H.; Seto, N. O. L.; Palcic, M. M.; Macher, B. A. Characterization of cysteine residues and disulfide bonds in proteins by liquid chromatography/electrospray ionization tandem mass spectrometry. *J. Mass Spectrom.* **2000**, *35*, 990-1002.

(142) Frand, A. R.; Cuzzo, J. W.; Kaiser, C. A. Pathways for protein disulphide bond formation. *Trends in Cell Biol.* **2000**, *10*, 203-210.

(143) Gao, Q.; Xue, S.; Doneanu, C. E.; Shaffer, S. A.; Goodlett, D.R.; Nelson, S. D. Pro-CrossLink. Software tool for protein cross-linking and mass spectrometry. *Anal. Chem.* **2006**, *78*, 2145-2149.

(144) Panchaud, A.; Singh, P.; Shaffer, S. A.; Goodlett, D.R. xComb: A cross-linked peptide database approach to protein-protein interaction analysis. *J. Proteome Res.* **2010**, *9*, 2508-2515.

(145) Schilling, B.; Row, R. H.; Gibson, B. W.; Guo, X.; Young, M. M. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 834-850.

(146) Wefing, S.; Schnaible, V.; Hoffmann, D. SearchXLinks. A program for the identification of disulfide bonds in proteins from mass spectra. *Anal. Chem.* **2006**, *78*, 1235-1241.

(147) Huang, S. Y.; Hsieh, Y. T.; Chen, C. H.; Chen, C. C.; Sung, W. C.; Chou, M. Y.; Chen, S. F. Automatic disulfide bond assignment using a(1) ion screening by mass spectrometry for structural characterization of protein pharmaceuticals. *Anal. Chem.* **2012**, *84*, 4900-4906.

(148) Na, S.; Paek, E.; Lee, C. CIFTER: Automated charge-state determination for peptide tandem mass spectra. *Anal. Chem.* **2008**, *80*, 1520-1528.

(149) Sharma, V.; Eng, J. K.; Feldman, S.; von Haller, P. D.; MacCoss, M. J.; Noble, W. S. Precursor charge state prediction for electron transfer dissociation tandem mass spectra. *J. Proteome Res.* **2010**, *9*, 5438-5444.

(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

dissociation spectra. *Anal. Chem.* **2008**, *80*, 376-386.

(151) Carvalho, P. C.; Cociorva, D.; Wong, C. C. L.; Carvalho, M. D. D.; Barbosa, V. C.; Yates, J. R. Charge prediction machine: Tool for inferring precursor charge states of electron transfer dissociation tandem mass spectra. *Anal. Chem.* **2009**, *81*, 1996-2003.

## CHAPTER 2

### COLLISION INDUCED DISSOCIATION BEHAVIOR OF N-LINKED GLYCOPEPTIDES

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

Woodin, *et al.* GlycoPep Grader: A web-based utility for assigning the composition of *N*-linked glycopeptides. *Anal. Chem.* **2012**, *84*, 4821-4829.

#### 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.<sup>1</sup> 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.<sup>2, 3, 4, 5, 6, 7, 8, 9</sup> 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.<sup>10, 7, 8, 9, 11, 12</sup> 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.<sup>13, 14, 15, 16, 17, 18</sup>

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.<sup>13, 16</sup> 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.<sup>16, 17</sup> 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<sup>1</sup> 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.<sup>19</sup>

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.<sup>20</sup> 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.<sup>20, 21, 22, 23, 24</sup> 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.<sup>20, 24, 25</sup> 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.<sup>16</sup>

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 (CH<sub>3</sub> OH) and HPLC grade acetonitrile (CH<sub>3</sub> CN) were purchased from Fisher Scientific (Fairlawn, NJ). Ammonium bicarbonate (NH<sub>4</sub>HCO<sub>3</sub>) 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 M .

**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 NH<sub>4</sub>HCO<sub>3</sub> (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.*<sup>26</sup> 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, N<sub>2</sub>, 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  $q_z$  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 % H<sub>2</sub>O + 0.1 % formic acid and the mobile phase for solvent B

consisted of 99.9 % CH<sub>3</sub>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](http://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.<sup>25</sup> 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.

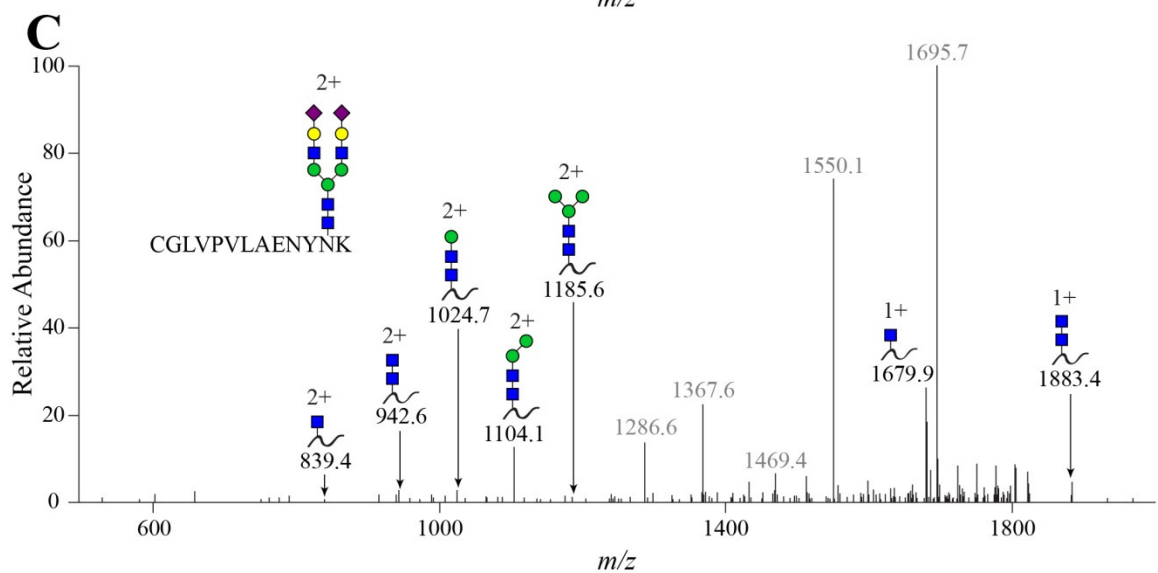
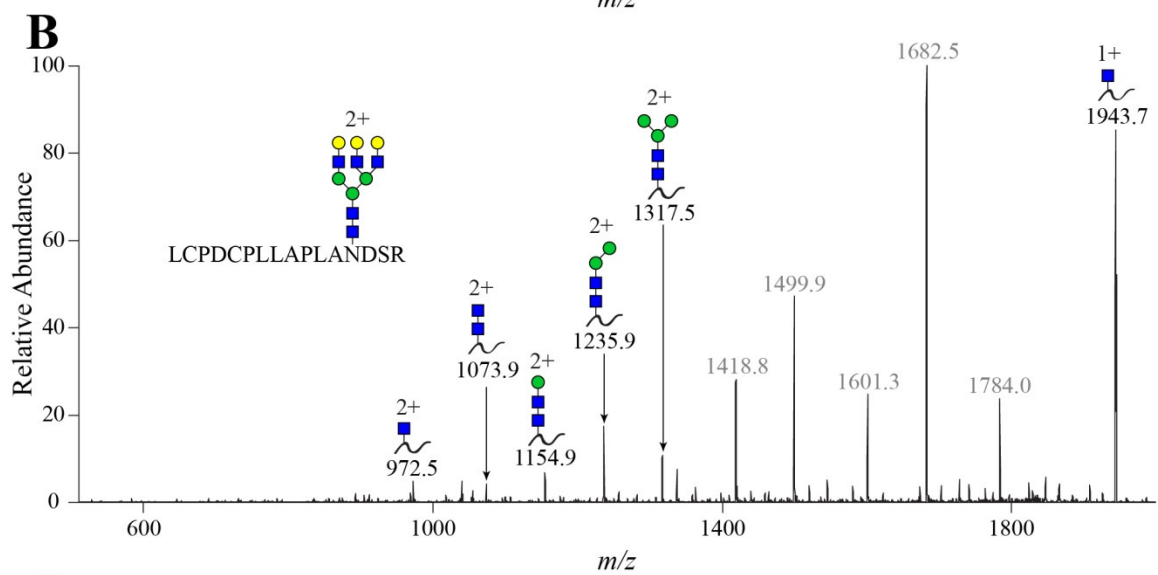
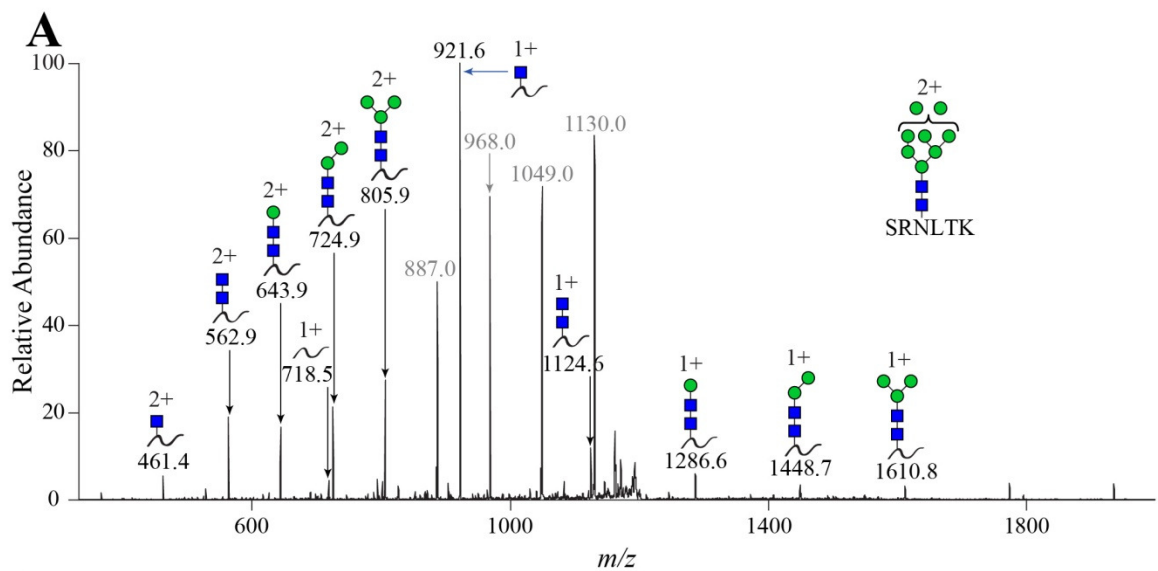
Glycoprotein	Mass (Da)	Length (AA)	# <i>N</i> -glycan Sites <sup>1</sup>	<i>N</i> -Glycan Type
RNase B	16,461	150	1	High Mannose
Asialofetuin	38,419	359	3	Complex
Transferrin	77,064	698	2	Sialylated Complex

<sup>1</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.

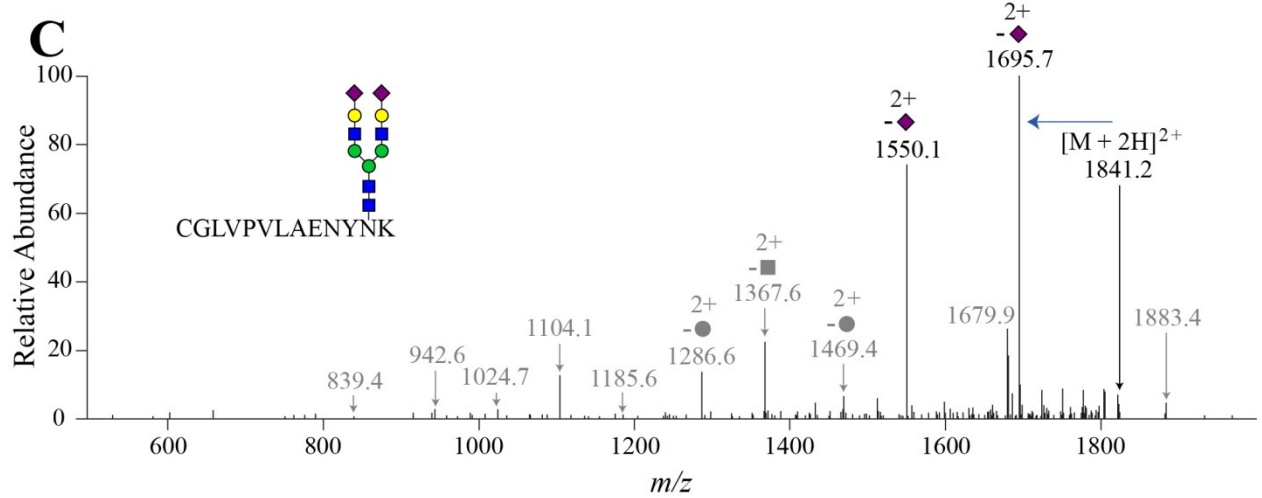
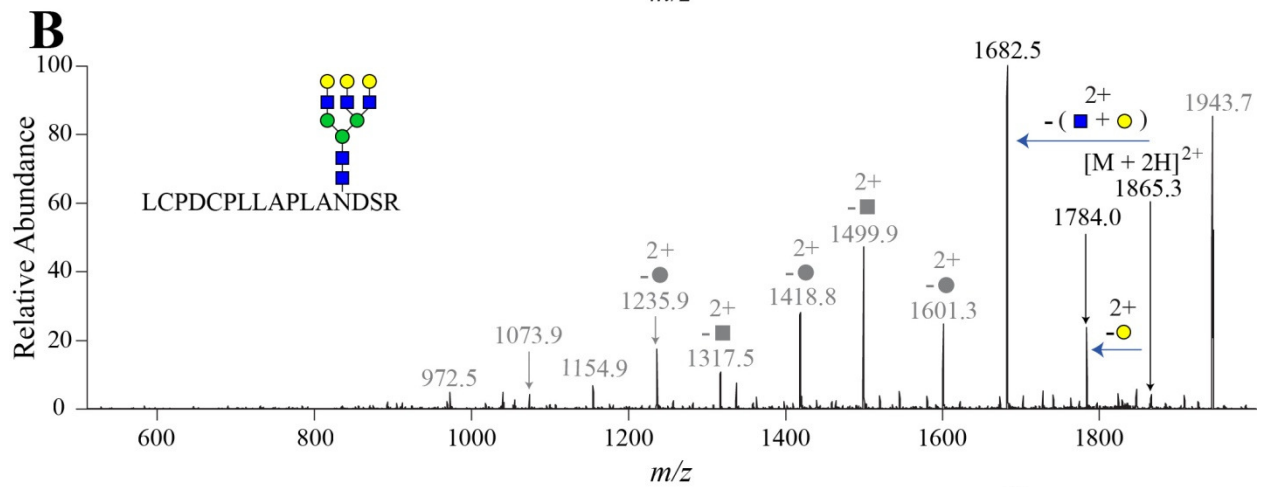
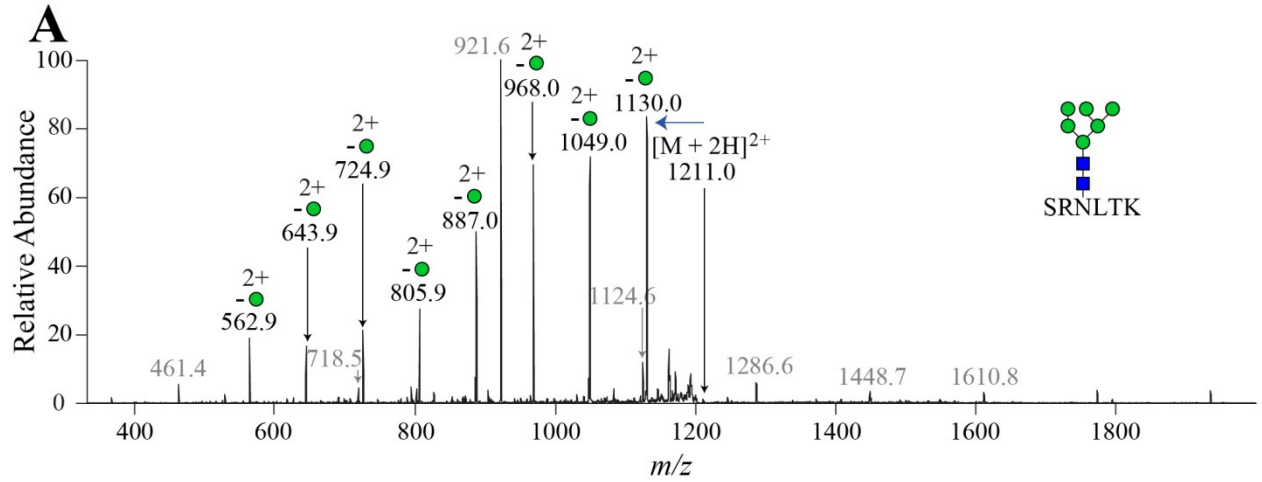
It has been shown previously that the same [peptide + core component] ions are present in CID spectra of glycopeptides.<sup>23</sup> From the MS/MS data we obtained, the [peptide + core component] product ions were also found to be present in multiple charge states, when the charge state of the precursor 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.<sup>27</sup>



**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.<sup>25, 28, 29</sup> 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.

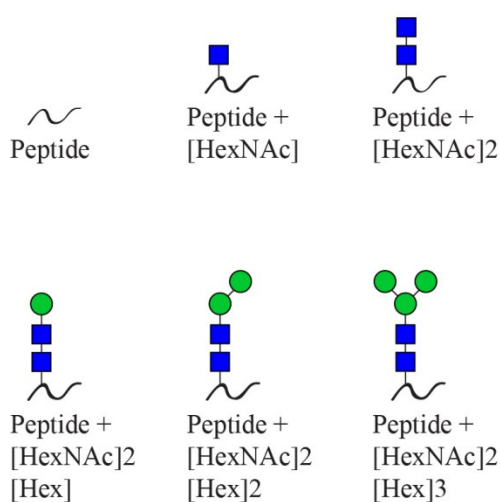
Although other neutral losses corresponding to the [precursor – monosaccharide] product ion types are often present in CID spectra collected on the glycopeptides, (these ions are in gray in the above figure) they were not shown to be unique enough to discriminate among various potential glycan substituent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

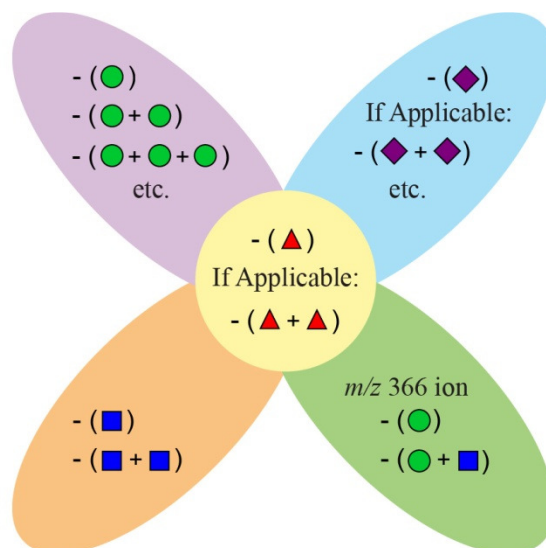
glycans, except no fucose is present); and 8) Complex/hybrid type structures that lack sialic acid or fucose and contain terminal Hex residues (which is the same as group 6 glycans, except no fucose is present).

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.<sup>25</sup> Figure 3 displays the product ions detected for each of the glycan class types devised on the basis of their fragmentation characteristics.

### A. [Peptide + Core Component] Ions



### B. [Precursor – Monosaccharide] Ions



**Figure 3.** 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.

**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:

$(\text{Spectra}[m/z] - 1) \leq \text{SearchCandidate} \leq (\text{Spectra}[m/z] + 1)$

and TotalRawPeptideScore and TotalRawGlycanScore are incremented only if:

$(\text{Spectra}[\text{Minima}]) \leq \text{SearchCandidate} \leq (\text{Spectra}[\text{Maxima}])$

#### **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:

$Y0 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073)) / \text{ChargeState}$

$Y1 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + \text{HexNAc}) / \text{ChargeState}$

$Y1 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + 203.0794) / \text{ChargeState}$

$Y2 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + \text{HexNAc} + \text{HexNAc}) / \text{ChargeState}$

$Y2 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + 203.0794 + 203.0794) / \text{ChargeState}$

$Y3 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + \text{HexNAc} + \text{HexNAc} + \text{Hex}) / \text{ChargeState}$

$Y3 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + 203.0794 + 203.0794 + 162.0528) / \text{ChargeState}$

$Y4 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + \text{HexNAc} + \text{HexNAc} + \text{Hex} + \text{Hex}) / \text{ChargeState}$

$Y4 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528) / \text{ChargeState}$

$Y5 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + \text{HexNAc} + \text{HexNAc} + \text{Hex} + \text{Hex} + \text{Hex}) / \text{ChargeState}$

$Y5 = (\text{PeptideMass} + (\text{ChargeState} * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528 + 162.0528) / \text{ChargeState}$

If match, ActualRawPeptideScore += 1. Regardless of match, TotalRawPeptideScore += 1.

2B. For ChargeStateIterator from DecrementChargeState to 1, recursively look for each of these candidates:

$Y0 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073)) / \text{ChargeStateIterator}$

$Y1 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + \text{HexNAc}) / \text{ChargeStateIterator}$

$Y1 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + 203.0794) / \text{ChargeStateIterator}$

$Y2 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + \text{HexNAc} + \text{HexNAc}) / \text{ChargeStateIterator}$

$Y2 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + 203.0794 + 203.0794) / \text{ChargeStateIterator}$

$Y3 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + \text{HexNAc} + \text{HexNAc} + \text{Hex}) / \text{ChargeStateIterator}$

$Y3 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + 203.0794 + 203.0794 + 162.0528) / \text{ChargeStateIterator}$

$Y4 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + \text{HexNAc} + \text{HexNAc} + \text{Hex} + \text{Hex}) / \text{ChargeStateIterator}$

$Y4 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528) / \text{ChargeStateIterator}$

$Y5 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + \text{HexNAc} + \text{HexNAc} + \text{Hex} + \text{Hex} + \text{Hex}) / \text{ChargeStateIterator}$

$Y5 = (\text{PeptideMass} + (\text{ChargeStateIterator} * 1.0073) + 203.0794 + 203.0794 + 162.0528 + 162.0528 + 162.0528) / \text{ChargeStateIterator}$

If match, ActualRawPeptideScore += 2. Regardless of match, TotalRawPeptideScore += 2.

2C. Look for the Y1 candidate at DecrementChargeState.

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 \* 146.0579) / 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 (Fuc + 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 - ((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:

$\text{SearchCandidate} = \text{PrecursorIon} - (((\text{CandidateFuc} * 146.0579) + (2 * 203.0794)) / \text{ChargeState})$   
 If match,  $\text{ActualRawGlycanScore} += 2$ . Regardless of match,  $\text{TotalRawGlycanScore} += 2$ .  
 D-1D-1F. Continue from G.  
 D-1D-2A. Look for loss of (Fuc + Hex) from PrecursorIon at ChargeState:  
 $\text{SearchCandidate} = \text{PrecursorIon} - ((146.0579 + 162.0528) / \text{ChargeState})$   
 If match,  $\text{ActualRawGlycanScore} += 4$ . Regardless of match,  $\text{TotalRawGlycanScore} += 4$ .  
 D-1D-2B. Is  $\text{CandidateFuc} > 1$ ? If yes, continue from D-1D-2C. If no, continue from D-1D-2D.  
 D-1D-2C. Look for loss of  $((\text{CandidateFuc} * \text{Fuc}) + \text{Hex})$  from PrecursorIon at ChargeState:  
 $\text{SearchCandidate} = \text{PrecursorIon} - (((\text{CandidateFuc} * 146.0579) + 162.0528) / \text{ChargeState})$   
 If match,  $\text{ActualRawGlycanScore} += 4$ . Regardless of match,  $\text{TotalRawGlycanScore} += 4$ .  
 D-1D-2D. Look for loss of (Fuc + Hex + HexNAc) from PrecursorIon at DecrementChargeState:  
 $\text{SearchCandidate} = ((\text{PrecursorIon} * \text{ChargeState}) - 1.0073 - 146.0579 - 162.0528 - 203.0794) / \text{DecrementChargeState}$   
 If match, continue from D-1D-2D-1A. If no match, continue from D-1D-2D-2A.  
 D-1D-2D-1A. Add 4 to  $\text{ActualRawGlycanScore}$ . Add 4 to  $\text{TotalRawGlycanScore}$ .  
 D-1D-2D-1B. Continue from F1.  
 D-1D-2D-2A. Look for loss of (Fuc + Hex + HexNAc) from PrecursorIon at ChargeState:  
 $\text{SearchCandidate} = \text{PrecursorIon} - ((146.0579 + 162.0528 + 203.0794) / \text{ChargeState})$   
 If match,  $\text{ActualRawGlycanScore} += 4$ . Regardless of match,  $\text{TotalRawGlycanScore} += 4$ .  
 D-1D-2D-2B. Continue from F1.  
 D-2. Is  $\text{CandidateNeu5Ac} > 0$ ? 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 DecrementChargeState:  
 $\text{SearchCandidate} = ((\text{PrecursorIon} * \text{ChargeState}) - 1.0073 - 203.0794) / \text{DecrementChargeState}$   
 If match,  $\text{ActualRawGlycanScore} += 4$ . Regardless of match,  $\text{TotalRawGlycanScore} += 4$ .  
 E-1B. Look for loss of (HexNAc) from PrecursorIon at ChargeState:  
 $\text{SearchCandidate} = \text{PrecursorIon} - (203.0794 / \text{ChargeState})$   
 If match,  $\text{ActualRawGlycanScore} += 4$ . Regardless of match,  $\text{TotalRawGlycanScore} += 4$ .  
 E-1C. Look for loss of  $(2 * \text{HexNAc})$  from PrecursorIon at ChargeState:  
 $\text{SearchCandidate} = \text{PrecursorIon} - ((2 * 203.0794) / \text{ChargeState})$   
 If match,  $\text{ActualRawGlycanScore} += 2$ . Regardless of match,  $\text{TotalRawGlycanScore} += 2$ .  
 E-1D. Continue from G.  
 E-2A. Look for loss of (Hex) from PrecursorIon at ChargeState:  
 $\text{SearchCandidate} = \text{PrecursorIon} - (162.0528 / \text{ChargeState})$   
 If match,  $\text{ActualRawGlycanScore} += 4$ . Regardless of match,  $\text{TotalRawGlycanScore} += 4$ .  
 E-2B. Look for loss of (Hex + HexNAc) from PrecursorIon at DecrementChargeState:  
 $\text{SearchCandidate} = ((\text{PrecursorIon} * \text{ChargeState}) - 1.0073 - 162.0528 - 203.0794) / \text{DecrementChargeState}$   
 If match, continue from E-2B-1A. If no match, continue from E-2B-2A.  
 E-2B-1A. Add 4 to  $\text{ActualRawGlycanScore}$ . Add 4 to  $\text{TotalRawGlycanScore}$ .  
 E-2B-1B. Continue from F1.  
 E-2B-2A. Look for loss of (Hex + HexNAc) from PrecursorIon at ChargeState:  
 $\text{SearchCandidate} = \text{PrecursorIon} - ((162.0528 + 203.0794) / \text{ChargeState})$   
 If match,  $\text{ActualRawGlycanScore} += 4$ . Regardless of match,  $\text{TotalRawGlycanScore} += 4$ .  
 E-2B-2B. Continue from F1.  
 F1. Look for Hex-HexNAc marker ion (366).  
 If match,  $\text{ActualRawGlycanScore} += 2$ . Regardless of match,  $\text{TotalRawGlycanScore} += 2$ .  
 3F2. Go to G.  
 3G.  $\text{GlycanScore} = \text{ActualRawGlycanScore} / \text{TotalRawGlycanScore}$   
 4.  $\text{GlycopeptideScore} = (\text{PeptideScore} * 0.67) + (\text{GlycanScore} * 0.33)$

## 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

- (1) Apweiler, R.; Hermjakob, H.; Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta-Gen. Subj.* **1999**, *1473*, 4-8.
- (2) Murrell, M. P.; Yarema, K. J.; Levchenko, A. The systems biology of glycosylation. *Chembiochem.* **2004**, *5*, 1334-1347.
- (3) Helenius, A.; Aebi, M. Intracellular functions of *N*-linked glycans. *Science.* **2001**, *291*, 2364-2369.
- (4) Petrescu, A. J.; Wormald, M. R.; Dwek, R. A. Structural aspects of glycomes with a focus on *N*-glycosylation and glycoprotein folding. *Curr. Opin. Struct. Biol.* **2006**, *16*, 600-607.
- (5) Skropeta, D. The effect of individual *N*-glycans on enzyme activity. *Bioorganic & Medicinal Chemistry.* **2009**, *17*, 2645-2653.
- (6) Bertozzi, C. R.; Kiessling, L. L. Chemical glycobiology. *Science.* **2001**, *291*, 2357-2364.
- (7) Spiro, R. G. Protein glycosylation: Nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology.* **2002**, *12*, 43R-56R.
- (8) Dennis, J. W.; Granovsky, M.; Warren, C. E. Protein glycosylation in development and disease. *BioEssays.* **1999**, *21*, 412-421.
- (9) Ohtsubo, K.; Marth, J. D. Glycosylation in cellular mechanisms of health and disease. *Cell.* **2006**, *126*, 855-867.
- (10) Drake P. M.; Cho, W.; Li, B.; Prakobphol, A.; Johansen, E.; Anderson, N. L.; Regnier, F.E.; Gibson, B.W.; Fisher S. J. Sweetening the pot: Adding glycosylation to the biomarker discovery equation. *Clin. Chem.* **2010**, *56*, 223-236.
- (11) Dube, D. H.; Bertozzi, C. R. Glycans in cancer and inflammation – Potential for therapeutics and diagnostics. *Nat. Rev. Drug Disc.* **2005**, *4*, 477-488.
- (12) Lefebvre, T.; Dehennaut, V.; Guinez, C.; Olivier, S.; Drougat, L.; Mir, A. M.; Mortuaire, M.; Vercoutter-Edouart, A. S.; Michalski, J. C. Dysregulation of the nutrient/stress sensor *O*-GlcNAcylation is involved in the etiology of cardiovascular disorders, type-2 diabetes and Alzheimer's disease. *Biochim. Biophys. Acta.-Gen. Subj.* **2010**, *1800*, 67-79.
- (13) Morelle, W.; Canis, K.; Chirat, F.; Faid, V.; Michalski, J. C. The use of mass spectrometry for the proteomic analysis of glycosylation. *Proteomics.* **2006**, *6*, 3993-4015.
- (14) Budnik, B. A., Lee, R. S.; Steen, J. A. J. Global methods for protein glycosylation analysis by mass spectrometry. *BBA-Protein Proteomics.* **2006**, *1764*, 1870-1880.

- (15) Zaia, J. Mass spectrometry and the emerging field of glycomics. *Chemistry & Biology*. **2008**, *15*, 881-892.
- (16) Dalpathado, D. S.; Desaire, H. Glycopeptide analysis by mass spectrometry. *Analyst*. **2008**, *133*, 731-738.
- (17) Schiel, J. E. Glycoprotein analysis using mass spectrometry: Unraveling the layers of complexity. *Anal. Bioanal. Chem.* **2012**, *404*, 1141-1149.
- (18) Morelle, W.; Michalski, J. C. The mass spectrometric analysis of glycoproteins and their glycan structures. *Curr. Anal. Chem.* **2005**, *1*, 29-57.
- (19) Desaire, H.; Hua, D. When can glycopeptides be assigned based solely on high-resolution mass spectrometry data? *Int. J. Mass. Spectrom.* **2009**, *287*, 21-26.
- (20) Huddleston, M. J.; Bean, M. F.; Carr, S. A. Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: Methods for selective detection of glycopeptides in protein digests. *Anal. Chem.* **1993**, *65*, 877-884.
- (21) Song, E.; Pyreddy, S.; Mechref, Y. Quantification of glycopeptides by multiple reaction monitoring liquid chromatography/tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 1941-1954.
- (22) Zaia, J. Mass spectrometry of oligosaccharides. *Mass Spectrom. Rev.* **2004**, *23*, 161-227.
- (23) Ritchie, M. A.; Gill, A. C.; Deery, M. J.; Lilley, K. Precursor ion scanning for detection and structural characterization of heterogeneous glycopeptide mixtures. *J. Am. Soc. Mass. Spectrom.* **2002**, *13*, 1065-1077.
- (24) Conboy, J. J.; Henion, J. D. The determination of glycopeptides by liquid-chromatography mass-spectrometry with collision-induced dissociation. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 804-814.
- (25) Nwosu, C. C.; Seipert, R. R.; Strum, J. S.; Hua, S. S.; An, H. J.; Zivkovic, A. M.; German, B. J.; Lebrilla, C. B. Simultaneous and extensive site-specific *N*- and *O*-glycosylation analysis in protein mixtures. *J. Proteome Res.* **2011**, *10*, 2612-2624.
- (26) Rebecchi, K.R.; Wenke, J. L.; Go, E.P.; Desaire, H. Label-free quantitation: A new glycoproteomics approach. *J. Am. Soc. Mass. Spectrom.* **2009**, *20*, 1048-1059.
- (27) Seipert, R. R.; Dodds, E. D.; Clowers, B. H.; Beecroft, S. M.; German, J. B.; Lebrilla, C. B. Factors that influence fragmentation behavior of *N*-linked glycopeptide ions. *Anal. Chem.* **2008**, *80*, 3684-3692.
- (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*

*Commun. Mass. Spectrom.* **2009**, 23, 161-170.

(29) Zhang, Z.; Shah, B. Prediction of collision-induced dissociation spectra of common *N*-linked glycopeptides for glycoform identification. *Anal. Chem.* **2010**, 82, 10194-10202.

## 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:**

Woodin, *et al.* GlycoPep Grader: A web-based utility for assigning the composition of *N*-linked glycopeptides. *Anal. Chem.* **2012**, *84*, 4821-4829.

#### 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 Y<sub>1</sub> 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.<sup>1, 2, 3, 4</sup> It is well-documented that glycosylation regulates a variety of intra- and extra-cellular processes.<sup>3, 4, 5, 6, 7, 8, 9</sup> Cellular communication and transport events,<sup>5, 6</sup> and mechanisms of protein folding,<sup>3, 5, 6</sup> degradation,<sup>3, 5</sup> and enzymatic interaction,<sup>7</sup> have all been shown to be regulated by glycosylation, the majority of which are *N*-linked in type.<sup>1</sup> 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.<sup>3, 8, 10, 11, 12</sup>

Typically, glycosylation analysis using mass spectrometry (MS) techniques is accomplished using one of two approaches: Glycan analysis and glycopeptide analysis.<sup>12</sup> The most information-rich of these methods is glycopeptide analysis, as glycosylation characteristics at individual sites of glycan attachment are readily identifiable.<sup>2, 12</sup> 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<sup>13</sup> and GlycoPep DB<sup>14</sup> calculate glycopeptide candidate compositions on the basis of mass information, as do a number of custom-generated databases.<sup>15, 16</sup> Unfortunately, a large amount of mass redundancy is typically encountered in glycopeptide analysis. Many different combinations of glycan composition + peptide composition are isobaric,<sup>15</sup> 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

process is laborious, time-consuming, and requires significant expertise.<sup>3, 4, 17</sup>

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.<sup>18, 19, 20, 21, 22.</sup> In terms of those tools that are publicly available for glycopeptide analysis, many have been designed to predominantly analyze the fragmentation of glycans.<sup>23, 24</sup> 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.<sup>23, 24</sup> GlycoWorkBench<sup>23</sup> and Glyco-Peakfinder<sup>24</sup> 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.*<sup>25</sup> 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.<sup>25</sup>

The most promising publicly accessible tool specifically developed to interpret and score MS/MS data of glycopeptides is GlycoMiner, developed by Ozohanics *et al.*<sup>26</sup> 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.<sup>26</sup> 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.<sup>26</sup>

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 as 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

stated.<sup>27, 28, 29</sup>

**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.<sup>27</sup> 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.<sup>27</sup> 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.<sup>30, 31, 32</sup> 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](http://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 gp140 CFI, which had been previously analyzed in our laboratory.<sup>27</sup> 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.<sup>27</sup>

**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<sup>13</sup> or GlycoPep DB<sup>14</sup>, or custom-generated databases.<sup>15, 16</sup> 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: [HexNAc]n[Hex]n[Neu5Ac]n[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 on-line tool, GlycoPep DB.<sup>14</sup>) 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.

Glycoprotein	Mass (Da)	Length (AA)	# <i>N</i> -glycan Sites	<i>N</i> -Glycan Type
RNase B <sup>1</sup>	16,461	150	1	High Mannose
Asialofetuin <sup>1</sup>	38,419	359	3	Complex
Transferrin <sup>1</sup>	77,064	698	2	Sialylated Complex
CON-S gp140 CFI <sup>2</sup>	~140,000	610	21	Highly Diverse

<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,<sup>33</sup> 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.

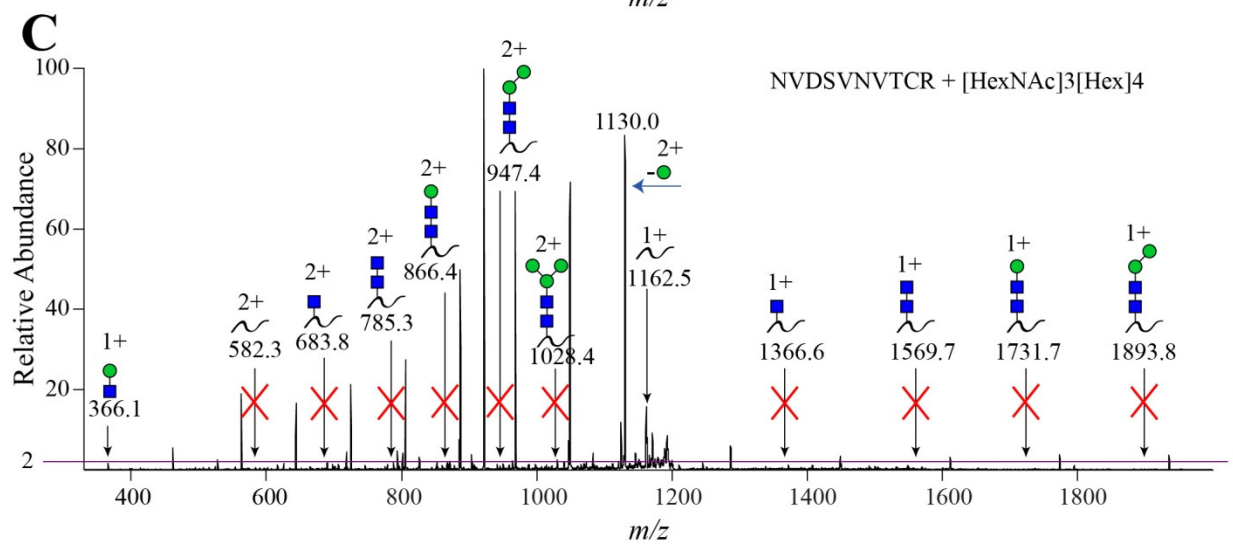
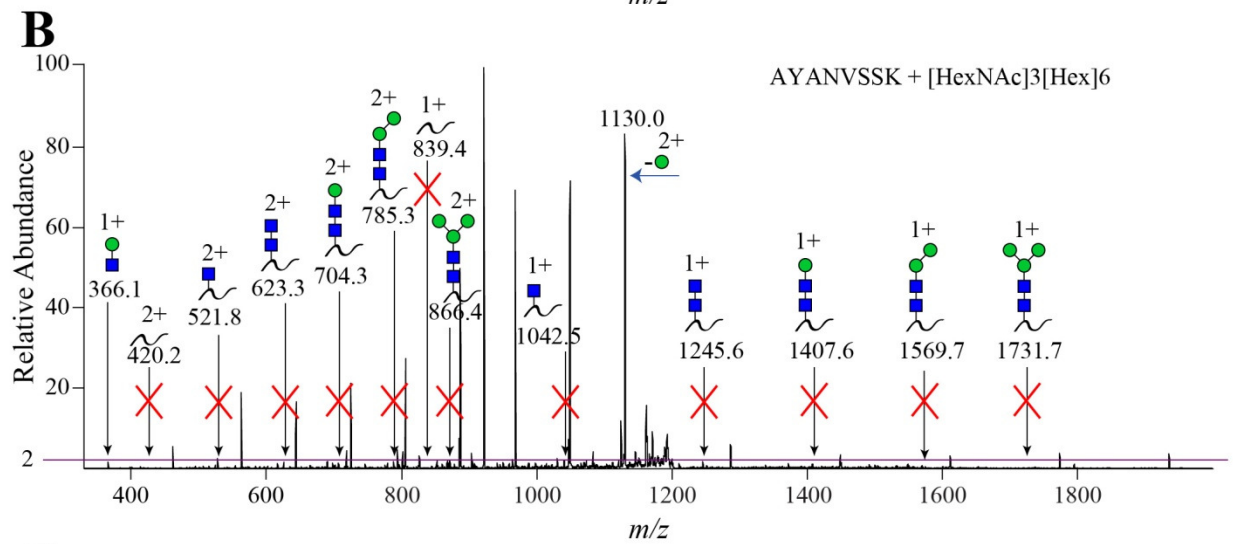
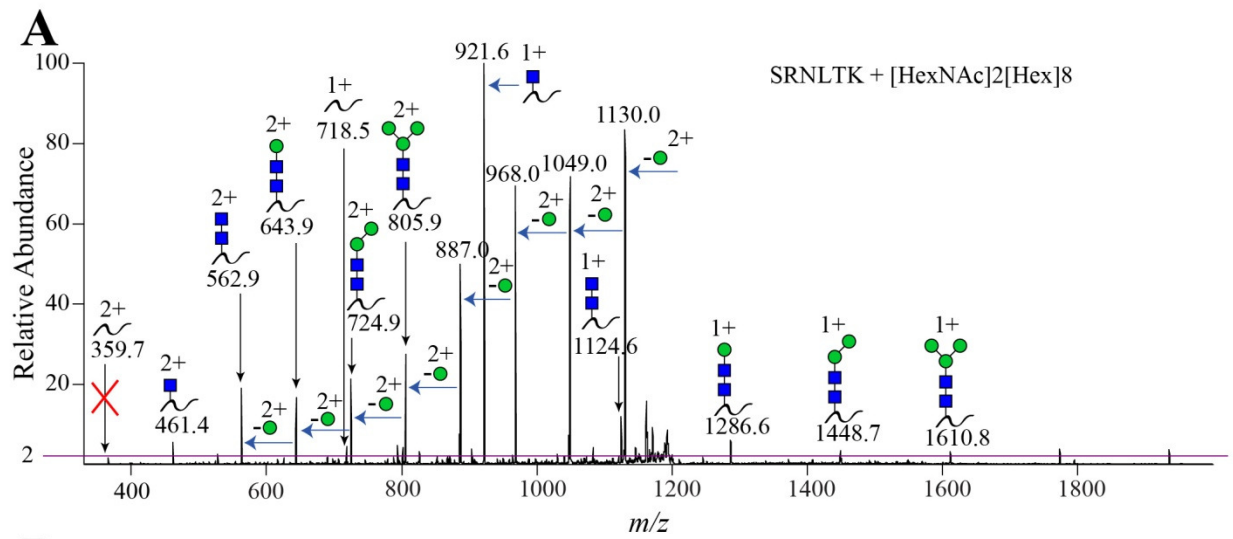
The relative abundance of the [precursor – monosaccharide] product ions is also taken into account when determining whether or not a peak corresponding to a particular  $m/z$  is actually from the neutral loss being evaluated, with varying threshold limits being applied according to the composition of the monosaccharide residues 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 MS<sup>2</sup> 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).





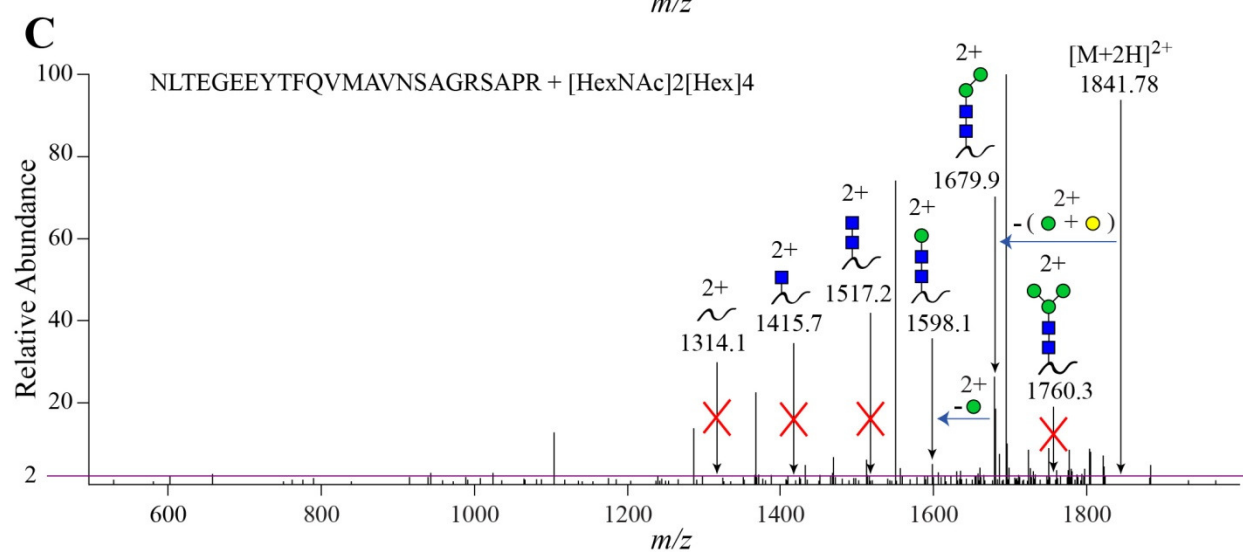
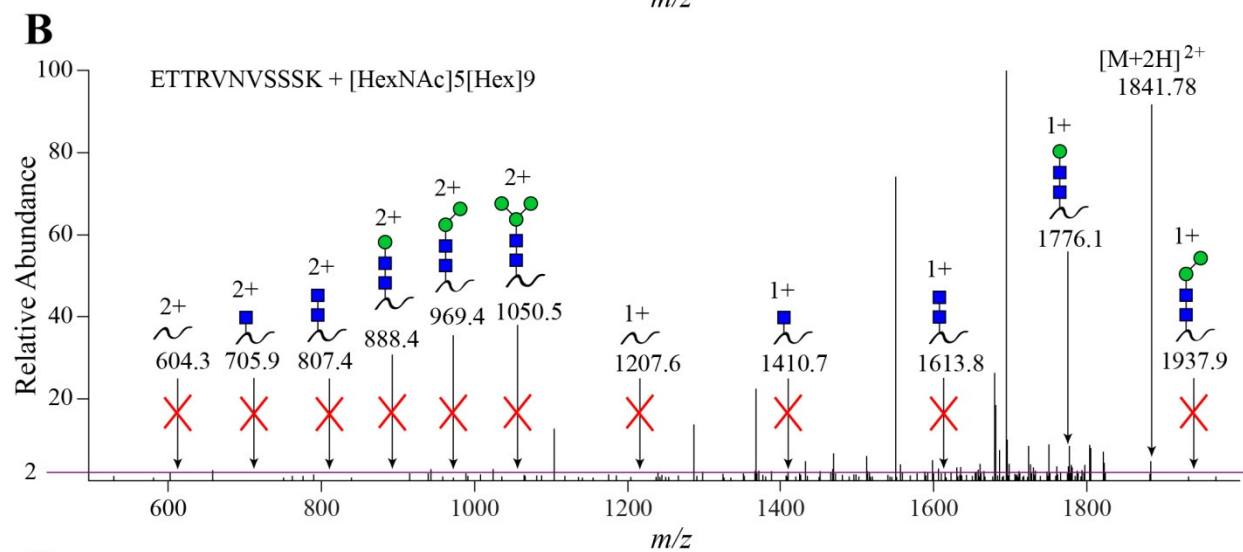
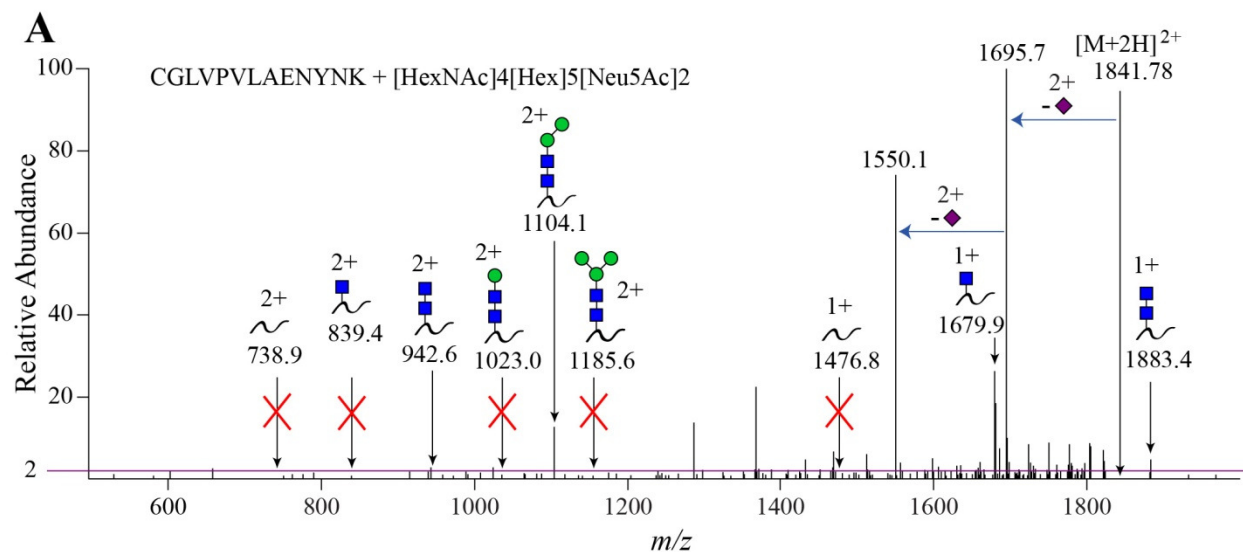
**Figure 1.** CID data of an RNase B glycopeptide from the training data set. (A) GPG scoring of the correct glycopeptide composition: 97 %. (B), (C) Scoring of two decoy compositions of the same nominal mass: 20 % and 27 %, respectively. Exact neutral masses of candidate compositions shown in (A), (B), 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 7 % and 39 % to the two decoy compositions of ETTRVNVSSSK + [HexNAc]5[Hex]9 and NLTEGEEYTFQVMAVNSAGRSAPR + [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 MS<sup>2</sup> 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.

Test <sup>1</sup>	Charge	Candidate <sup>2</sup>	<i>m/z</i>	Glycopeptide Composition	GPG <sup>3</sup>
1	2+	A	967.9252	SRNLTK + [HexNAc]2[Hex]5	91
	2+	B	967.8831	NASHK + [HexNAc]2[Hex]6	95
	2+	C	967.9336	WVRHNK + [HexNAc]3[Hex]3	20
2	2+	A	981.9227	NLTKDR + [HexNAc]2[Hex]5	91
	2+	B	981.9125	NRSLTN + [HexNAc]3[Hex]4	26
	2+	C	981.9550	RETQAVNWTK + [HexNAc]2[Hex]2	0
3	2+	A	1048.9516	SRNLTK + [HexNAc]2[Hex]6	97
	2+	B	1048.9095	NASHK + [HexNAc]2[Hex]7	63
	2+	C	1048.9600	WVRHNK + [HexNAc]3[Hex]4	20
	2+	D	1048.9492	NVDSVVGTCR <sup>4</sup> + [HexNAc]3[Hex]2	23
4	2+	A	1103.4892	SRNLTKDR + [HexNAc]2[Hex]5	63
	2+	B	1103.6655	NRSLTN + [HexNAc]5[Hex]3	0
	2+	C	1103.4863	PFNQTKNRF + [HexNAc]2[Hex]4	38
5	2+	A	1129.9780	SRNLTK + [HexNAc]2[Hex]7	91
	2+	B	1129.9756	NVDSVVGTCR <sup>4</sup> + [HexNAc]3[Hex]3	40
	2+	C	1129.9674	AYANVSSK + [HexNAc]3[Hex]5	44
	2+	D	1129.9864	WVRHNK + [HexNAc]3[Hex]5	44
6	2+	A	1143.9755	NLTKDR + [HexNAc]2[Hex]7	97
	2+	B	1144.0078	RETQAVNWTK + [HexNAc]2[Hex]4	48
	2+	C	1144.0129	NVTGTTSETIK + [HexNAc]4[Hex]2	30
	2+	D	1143.9516	ANK + [HexNAc]5[Hex]4[Neu5Ac]1	27
7	2+	A	1184.5156	SRNLTKDR + [HexNAc]2[Hex]6	63
	2+	B	1184.5127	PFNQTKNRF + [HexNAc]2[Hex]5	47
	2+	C	1184.5015	VNVSSSK + [HexNAc]5[Hex]3[Fuc]1	0
8	2+	A	1211.0044	SRNLTK + [HexNAc]2[Hex]8	97
	2+	B	1211.0128	WVRHNK + [HexNAc]3[Hex]6	38
	2+	C	1211.0020	NVDSVVGTCR <sup>4</sup> + [HexNAc]3[Hex]4	33
9	2+	A	1225.0019	NLTKDR + [HexNAc]2[Hex]8	94
	2+	B	1225.0207	NLTKDRGP + [HexNAc]3[Hex]4[Neu5Ac]1	21
	2+	C	1225.0393	NVTGTTSETIK + [HexNAc]4[Hex]3	10
	2+	D	1225.0342	RETQAVNWTK + [HexNAc]2[Hex]5	47
10	2+	A	1265.5420	SRNLTKDR + [HexNAc]2[Hex]7	82
	2+	B	1265.5183	NRSLTN + [HexNAc]5[Hex]5	20
	2+	C	1265.5309	VNVSSSK + [HexNAc]5[Hex]4[Fuc]1	25
11	2+	A	1292.0308	SRNLTK + [HexNAc]2[Hex]9	91
	2+	B	1292.0284	NVDSVVGTCR <sup>4</sup> + [HexNAc]3[Hex]5	47
	2+	C	1292.0392	AYANVSSK + [HexNAc]3[Hex]7	38
	2+	D	1292.0392	WVRHNK + [HexNAc]3[Hex]7	38
12	2+	A	1306.0283	NLTKDR + [HexNAc]2[Hex]9	88
	2+	B	1306.0654	NLTKDRL + [HexNAc]4[Hex]4[Neu5Ac]1	24
	2+	C	1306.0781	LVINR + [HexNAc]6[Hex]3[Fuc]2	19
	2+	D	1306.0606	RETQAVNWTK + [HexNAc]2[Hex]6	57
13	2+	A	1346.5684	SRNLTKDR + [HexNAc]2[Hex]8	76

	2 +	B	1346.5447	NRSLTN + [HexNAc]5[Hex]6	23
	2 +	C	1346.6205	LNVTLKWTK + [HexNAc]4[Hex]3[Neu5Ac]1	12
	2 +	D	1346.5913	DNGSPILGYWLEK + [HexNAc]2[Hex]4[Fuc]1	4
14	5 +	A	1059.6808	RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR + [HexNAc]4[Hex]5	51
	5 +	B	1059.6980	IENNTTVLKSSATFQSTVAGSPISITWLK + [HexNAc]5[Hex]5[Neu5Ac]1	12
	5 +	C	1059.6567	CHYMTIHNVTDPDEGVYSVIARLEPR <sup>4</sup> + [HexNAc]5[Hex]6[Neu5Ac]1	12
	5 +	D	1059.6535	NAAGNFSEPSDSSGAITARDEIDAPNASLDPK + [HexNAc]4[Hex]6[Fuc]2	17
15	5 +	A	1132.7073	RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR + [HexNAc]5[Hex]6	61
	5 +	B	1132.6652	SCEPVPARDPCDPPGQPEVTNITR <sup>4</sup> + [HexNAc]6[Hex]7[Neu5Ac]2[Fuc]1	25
	5 +	C	1132.7244	IENNTTVLKSSATFQSTVAGSPISITWLK + [HexNAc]6[Hex]6[Neu5Ac]1	31
	5 +	D	1132.6708	HILVINDSQFDDEGVYTAEEVGK + [HexNAc]6[Hex]7[Neu5Ac]2[Fuc]1	25
16	4 +	A	1160.5442	VVHAVEVALATFNAESNGSYLQLVEISR + [HexNAc]4[Hex]5	67
	4 +	B	1160.5187	TDTMRLLERPPEFTLPLYNK + [HexNAc]4[Hex]5[Neu5Ac]2	29
	4 +	C	1160.5254	NVTVIEGESVTLECHISGYSPVTWYR <sup>4</sup> + [HexNAc]5[Hex]3	39
	4 +	D	1160.5186	LTPESTREFLCINGSIHQPLK <sup>4</sup> + [HexNAc]4[Hex]8	57
17	3 +	A	1164.5188	KLCPDCPLLAPLNSDR + [HexNAc]4[Hex]5	48
	3 +	B	1164.5166	NLNVRYQSNATLVCK <sup>4</sup> + [HexNAc]4[Hex]5[Fuc]1	9
	3 +	C	1164.4872	NSVGKSNCTVSVHVS <sup>4</sup> + [HexNAc]2[Hex]8	5
	3 +	D	1164.5183	VNKSLLNALK <sup>4</sup> + [HexNAc]5[Hex]4[Neu5Ac]2[Fuc]1	6
18	3 +	A	1243.5312	LCPDCPLLAPLNSDR + [HexNAc]5[Hex]6	59
	3 +	B	1243.5766	DVTALENATVAFEVSVSHDTVPVK + [HexNAc]2[Hex]4[Fuc]1	45
	3 +	C	1243.5153	YDSGKYTLLENSSGTK + [HexNAc]2[Hex]9	38
	3 +	D	1243.5069	RETQAVNWTK + [HexNAc]4[Hex]5[Neu5Ac]3	13
19	3 +	A	1286.2296	KLCPDCPLLAPLNSDR + [HexNAc]5[Hex]6	31
	3 +	B	1286.2375	EFLCINGSIHQPLK <sup>4</sup> + [HexNAc]8[Hex]3	5
	3 +	C	1286.1893	DSVNLTWTEPASDGGGSK + [HexNAc]4[Hex]7[Fuc]1	4
	3 +	D	1286.1844	KAYATITNNCTK <sup>4</sup> + [HexNAc]4[Hex]7[Neu5Ac]2	0
20	4 +	A	1324.3492	RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR + [HexNAc]4[Hex]5	70
	4 +	B	1324.3458	TLKNLTVTETQDAVFTVELTHPNVK + [HexNAc]4[Hex]5[Neu5Ac]3	16
	4 +	C	1324.3157	DSVNLTWTEPASDGGGSKITNYIVEK + [HexNAc]5[Hex]6[Neu5Ac]2	16
	4 +	D	1324.3101	LPYTTGPPSTPWVTNVTR + [HexNAc]6[Hex]6[Neu5Ac]3[Fuc]1	25
21	4 +	A	1415.6259	RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR + [HexNAc]5[Hex]6	37
	4 +	B	1415.5982	NAAGNFSEPSDSSGAITARDEIDAPNASLDPK + [HexNAc]5[Hex]7[Fuc]2	0
	4 +	C	1415.6288	TLKNLTVTETQDAVFTVELTHPNVK + [HexNAc]5[Hex]6[Neu5Ac]3	0
22	3 +	A	1547.0565	VVHAVEVALATFNAESNGSYLQLVEISR +	94

				[HexNAc]4[Hex]5	
	3 +	B	1547.0224	TDTMRLLERPPEFTLPLYNK + [HexNAc]4[Hex]5[Neu5Ac]2	40
	3 +	C	1547.0223	LTPESTREFLCINGSIHFAQPLK <sup>4</sup> + [HexNAc]4[Hex]8	67
	3 +	D	1547.0314	NVTVIEGESVTLECHISGYPSPTVTWYR <sup>4</sup> + [HexNAc]5[Hex]3	52
23	3 +	A	1668.7672	VVHAVEVALATFNAESNGSYLQLVEISR + [HexNAc]5[Hex]6	85
	3 +	B	1668.7332	TDTMRLLERPPEFTLPLYNK + [HexNAc]5[Hex]6[Neu5Ac]2	21
	3 +	C	1668.7370	LTPESTREFLCINSIHFAQPLK + [HexNAc]5[Hex]9	57
	3 +	D	1668.7384	AMKDGVDHIPEDAQLETAENSSVIIIPECK <sup>4</sup> + [HexNAc]4[Hex]4[Neu5Ac]1	7
	3 +	E	1668.7950	ATAVVEVNVLDKPGPPAAFDITDVTNESCLLTWNPP <sup>4</sup> + [HexNAc]2[Hex]4	50
24	2 +	A	1682.2271	LCPDCPLLAPLNDSR + [HexNAc]4[Hex]5	95
	2 +	B	1682.1743	SNCTVSVHVSDR <sup>4</sup> + [HexNAc]4[Hex]5[Neu5Ac]1[Fuc]1	12
	2 +	C	1682.2459	LVINRTHASDEGPYK + [HexNAc]5[Hex]4	0
	2 +	D	1682.1895	VTNVTK + [HexNAc]7[Hex]7[Fuc]1	0
25	2 +	A	1864.7932	LCPDCPLLAPLNDSR + [HexNAc]5[Hex]6	95
	2 +	B	1864.7574	DSGYSLTAENSSGTDQK + [HexNAc]6[Hex]3	0
	2 +	C	1864.7289	AYATITNNCTK <sup>4</sup> + [HexNAc]4[Hex]7[Neu5Ac]2	0
	2 +	D	1864.7693	YDSGKYTLTENSSGTK + [HexNAc]2[Hex]9	7
26	4 +	A	848.3658	CGLVPVLAENYNK + [HexNAc]4[Hex]5[Neu5Ac]1	75
	4 +	B	848.3761	VNKTIHDTQFK + [HexNAc]4[Hex]7	61
	4 +	C	848.3853	IRDAHLDDQANYNVSLTNHR + [HexNAc]2[Hex]3[Fuc]1	31
27	4 +	A	921.1397	CGLVPVLAENYNK + [HexNAc]4[Hex]5[Neu5Ac]2	52
	4 +	B	921.1512	MSDAGKYTVVAGGNVSTAK + [HexNAc]5[Hex]5	38
	4 +	C	921.1133	DGFNITTSEK + [HexNAc]5[Hex]6[Neu5Ac]2	37
	4 +	D	921.1763	LLTQNSENITIENEHYTHLVMK + [HexNAc]2[Hex]4	30
28	5 +	A	944.7902	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]4[Hex]5[Neu5Ac]2	36
	5 +	B	944.8100	GQVDLVDTMAFLVIPNSTR + [HexNAc]6[Hex]7[Neu5Ac]1	26
	5 +	C	944.7862	NNTLVLQVR + [HexNAc]6[Hex]7[Neu5Ac]4[Fuc]1	10
	5 +	D	944.8017	NVTFTSVIRGTPPFK + [HexNAc]5[Hex]9[Neu5Ac]2	15
29	4 +	A	1107.9621	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]4[Hex]5[Neu5Ac]1	46
	4 +	B	1107.9661	VNRLNVTLK + [HexNAc]6[Hex]7[Neu5Ac]3[Fuc]1	13
	4 +	C	1107.9878	ANDTLVRSTEYPCAGLVEGLEYSFR + [HexNAc]3[Hex]6	36
	4 +	D	1107.9765	NVTFTSVIRGTPPFK + [HexNAc]5[Hex]9[Neu5Ac]1	22
	4 +	E	1107.9609	NSILWTKVNK + [HexNAc]6[Hex]7[Neu5Ac]3	20
30	4 +	A	1180.7359	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]4[Hex]5[Neu5Ac]2	42
	4 +	B	1180.7302	RANHTPESCPETKYK + [HexNAc]6[Hex]5[Neu5Ac]3	12
	4 +	C	1180.7608	HILVINDSQFDEGVYTAEEVGK + [HexNAc]6[Hex]3[Neu5Ac]1[Fuc]1	8
	4 +	D	1180.7617	ANDTLVRSTEYPCAGLVEGLEYSFR + [HexNAc]3[Hex]6[Neu5Ac]1	15
31	3 +	A	1227.8504	CGLVPVLAENYNK + [HexNAc]4[Hex]5[Neu5Ac]2	35
	3 +	B	1227.8153	DGFNITTSEK + [HexNAc]5[Hex]6[Neu5Ac]2	19
	3 +	C	1227.8591	QNATVQGLIQGK + [HexNAc]5[Hex]6[Neu5Ac]1[Fuc]1	4

	3 +	D	1227.8656	MSDAGKYTVVAGGNVSTAK + [HexNAc]5[Hex]5	12
	3 +	E	1227.9028	ATMRFNTEITAENLTINLK + [HexNAc]5[Hex]3	18
32	3 +	A	1379.9152	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]4[Hex]5	76
	3 +	B	1379.9344	NVTFTSVIRGTPPFK + [HexNAc]5[Hex]9	38
	3 +	C	1379.9085	NNTLVLQVR + [HexNAc]6[Hex]7[Neu5Ac]2[Fuc]1	8
	3 +	D	1379.8840	VHTNATIR + [HexNAc]6[Hex]7[Neu5Ac]3	8
33	4 +	A	1434.3702	CGLVPVLAENYNKSDNCEDTPEAGYFAVAVVK + [HexNAc]4[Hex]5[Neu5Ac]2	53
	4 +	B	1434.3555	LNGSAPIQVCWYRDGVLLR + [HexNAc]6[Hex]7[Neu5Ac]4	25
	4 +	C	1434.4053	YTVTATNSAGTATENLSVIVLEKPGPPVGPVR + [HexNAc]4[Hex]5[Neu5Ac]3	19
	4 +	D	1434.3639	DNKEIRPGGNYTITCVGNTPHLR + [HexNAc]7[Hex]6[Neu5Ac]2[Fuc]1	14
34	3 +	A	1476.9470	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]4[Hex]5[Neu5Ac]1	82
	3 +	B	1476.9160	VHTNATIR + [HexNAc]6[Hex]7[Neu5Ac]4	8
	3 +	C	1476.9793	HILVINDSQDDEGVYTAEEVGK + [HexNAc]6[Hex]3[Fuc]1	13
	3 +	D	1476.9800	GQVDLVDTMAFLVIPNSTR + [HexNAc]6[Hex]7	6
35	2 +	A	1513.1582	CGLVPVLAENYNK + [HexNAc]3[Hex]4[Neu5Ac]1	96
	2 +	B	1513.1788	VNKTIHDTQFK + [HexNAc]3[Hex]6	0
	2 +	C	1513.1711	QNATVQGLIQGK + [HexNAc]4[Hex]5[Fuc]1	13
	2 +	D	1513.1284	NNVTLK + [HexNAc]6[Hex]6[Fuc]1	0
36	3 +	A	1525.6329	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]4[Hex]5[Neu5Ac]1[Fuc]1	52
	3 +	B	1525.6292	INNLTESDQGEYVCEISGEGGTSK + [HexNAc]5[Hex]6	0
	3 +	C	1525.6659	GQVDLVDTMAFLVIPNSTR + [HexNAc]6[Hex]7[Fuc]1	0
	3 +	D	1525.6018	VHTNATIR + [HexNAc]6[Hex]7[Neu5Ac]4[Fuc]1	11
37	3 +	A	1574.0003	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]4[Hex]5[Neu5Ac]2	76
	3 +	B	1574.0118	GQVDLVDTMAFLVIPNSTR + [HexNAc]6[Hex]7[Neu5Ac]1	33
	3 +	C	1573.9771	NSILWTKVVK + [HexNAc]6[Hex]7[Neu5Ac]4	17
	3 +	D	1574.0243	LLQSENITIENTEHYTHLVMK + [HexNAc]4[Hex]7[Fuc]1	0
38	3 +	A	1647.3437	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]5[Hex]6[Neu5Ac]1[Fuc]1	76
	3 +	B	1647.3598	CGPGEPAYVDEPVNMSTPATVPDPENVK + [HexNAc]3[Hex]6[Neu5Ac]1	48
	3 +	C	1647.3421	NSILWTKVVK + [HexNAc]7[Hex]8[Neu5Ac]3[Fuc]1	20
	3 +	D	1647.4079	AWTPVTYTVTRQNATVQGLIQGK + [HexNAc]5[Hex]5[Neu5Ac]2	23
39	2 +	A	1695.7243	CGLVPVLAENYNK + [HexNAc]4[Hex]5[Neu5Ac]1	71
	2 +	B	1695.7079	YILTVENSSGSK + [HexNAc]4[Hex]7[Fuc]1	0
	2 +	C	1695.6714	DGFNITTSEK + [HexNAc]5[Hex]6[Neu5Ac]1	43
	2 +	D	1695.7343	ANKTPIRM + [HexNAc]6[Hex]4[Neu5Ac]1[Fuc]1	11
	2 +	E	1695.7485	DGQTLKETTRVNVSSSK + [HexNAc]2[Hex]7	0
40	3 +	A	1695.6895	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]5[Hex]6[Neu5Ac]2	45
	3 +	B	1695.6829	NNTLVLQVR + [HexNAc]4[Hex]8[Fuc]1	15
	3 +	C	1695.7014	MSDAGKYTVVAGGNVSTAK + [HexNAc]8[Hex]9[Fuc]1	0
	3 +	D	1695.7225	WVRCNFTDVSECQYTVTGLSPGDR + [HexNAc]4[Hex]6[Neu5Ac]1[Fuc]1	39
41	3 +	A	1792.7213	QQQHLFGSNVTDCSGNFCLFR +	57



				[HexNAc]5[Hex]6[Neu5Ac]3	
	3 +	B	1792.7333	WVRCNFTDVSECQYTVTGLSPGDR + [HexNAc]4[Hex]7[Neu5Ac]2	39
	3 +	C	1792.7783	CHYMTIHNVTDPDEGVYSVIARLEPR + [HexNAc]6[Hex]4[Neu5Ac]1[Fuc]1	11
	3 +	D	1792.7840	LNWTKPEHDDGGAKIESYVIEMLK + [HexNAc]7[Hex]8	12
42	3 +	A	1815.1260	CGLVPVLAENYNKSDNCEDTPEAGYFAVAVVK + [HexNAc]4[Hex]5[Neu5Ac]1	92
	3 +	B	1815.1366	VSDVSRDSVNLWTWTEPASDGGGSKITNYIVEK + [HexNAc]4[Hex]6[Neu5Ac]1	52
	3 +	C	1815.1624	LLQSENIENTEHEHYTHLVMKNVQRK + [HexNAc]6[Hex]6	4
	3 +	D	1815.1173	YTLTVKNASGTKAVSVMVK + [HexNAc]7[Hex]8[Neu5Ac]2[Fuc]1	8
43	2 +	A	1841.2720	CGLVPVLAENYNK + [HexNAc]4[Hex]5[Neu5Ac]2	75
	2 +	B	1841.2534	ETTRVNVSSSK + [HexNAc]5[Hex]9	7
	2 +	C	1841.3143	NLTEGEEYTFQVMAVNSAGRSAPR + [HexNAc]2[Hex]4	39
	2 +	D	1841.3346	TEIISTDNHTLLTVK + [HexNAc]6[Hex]3[Fuc]2	7
44	2 +	A	1886.8030	QQQHLFGSNVTDCSGNFCLFR + [HexNAc]3[Hex]4	78
	2 +	B	1886.8111	VNRLNVTLK + [HexNAc]5[Hex]6[Neu5Ac]2[Fuc]1	6
	2 +	C	1886.8318	NVTFTSVIRGTPPFK + [HexNAc]4[Hex]8	66
	2 +	D	1886.8006	NSILWTKVVK + [HexNAc]5[Hex]6[Neu5Ac]2	0
45	3 +	A	1912.1578	CGLVPVLAENYNKSDNCEDTPEAGYFAVAVVK + [HexNAc]4[Hex]5[Neu5Ac]2	71
	3 +	B	1912.1382	LNGSAPIQVCWYRDGVLLR + [HexNAc]6[Hex]7[Neu5Ac]4	17
	3 +	C	1912.1494	DNKEIRPGGNYTITCVGNTPHLR + [HexNAc]7[Hex]6[Neu5Ac]2[Fuc]1	17
	3 +	D	1912.1546	TDTMRLLERPPEFTLPLYNK + [HexNAc]7[Hex]8[Neu5Ac]2	39

<sup>1</sup> For each test, candidate A is the actual glycopeptide composition that corresponds to the MS/MS data being scored.

<sup>2</sup> All candidate compositions have an  $m/z$  value, calculated *in silico*, within 50 ppm error of the monoisotopic mass present in the experimental MS<sup>1</sup> data.

<sup>3</sup> 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<sub>1</sub> 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.

<sup>4</sup> 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.

Test <sup>1</sup>	Charge	Candidate <sup>2</sup>	<i>m/z</i>	Glycopeptide Composition	GPG <sup>3</sup>
1	2 +	A	1071.5123	SNITGLLLTR + [HexNAc]2[Hex]4	83
	2 +	B	1071.4758	ANVTVEAR + [HexNAc]4[Hex]2	0
	2 +	C	1071.4695	MANISRYYEAPP + [HexNAc]2[Hex]2	8
2	3 +	A	1156.4722	DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]5	91
	3 +	B	1156.5238	YTLTVENNSGSKSITFTVK + [HexNAc]2[Hex]6	55
	3 +	C	1156.4895	VDQHEWTKCNTTPTK + [HexNAc]4[Hex]5	25
	3 +	D	1156.5225	NNLPISISSNVSISR + [HexNAc]6[Hex]4	0
3	2 +	A	1160.4919	EANTTLFCASDAK + [HexNAc]2[Hex]3	83
	2 +	B	1160.4997	ANVTVEAR + [HexNAc]4[Hex]4	31
	2 +	C	1160.5104	YQSNATLVCK + [HexNAc]4[Hex]2	25
	2 +	D	1160.4567	ANK + [HexNAc]5[Hex]6	25
4	2 +	A	1193.5653	SNITGLLLTR + [HexNAc]4[Hex]3	94
	2 +	B	1193.5372	LNWTKPEHDGGAK + [HexNAc]3[Hex]2	45
	2 +	C	1193.5239	YQSNATLPGKGN + [HexNAc]4[Hex]2	50
5	2 +	A	1254.0784	SNITGLLLTR + [HexNAc]3[Hex]5	97
	2 +	B	1254.0502	LNWTKPEHDGGAK + [HexNAc]2[Hex]4	58
	2 +	C	1254.0503	ANVTVEAR + [HexNAc]5[Hex]3[Fuc]1	0
	2 +	D	1254.0399	YCVVVENSTGSR + [HexNAc]4[Hex]2	33
6	3 +	A	1264.5074	DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]7	97
	3 +	B	1264.5590	YTLTVENNSGSKSITFTVK + [HexNAc]2[Hex]8	56
	3 +	C	1264.5247	VDQHEWTKCNTTPTK + [HexNAc]4[Hex]7	30
	3 +	D	1264.5171	NASGSAKAEIK + [HexNAc]5[Hex]6[Neu5Ac]2[Fuc]1	0
7	3 +	A	1268.8550	NNNNTNDTITLPCR + [HexNAc]6[Hex]4[Neu5Ac]1	53
	3 +	B	1268.8474	YTCQAKNESGVER + [HexNAc]5[Hex]5[Neu5Ac]1[Fuc]1	17
	3 +	C	1268.8582	QSDAGEYTFVAGRNR + [HexNAc]5[Hex]6[Fuc]1	5
	3 +	D	1268.8608	TKANVTVEAR + [HexNAc]5[Hex]6[Neu5Ac]2[Fuc]1	22
	3 +	E	1268.8474	YTCQAKNESGVER + [HexNAc]5[Hex]5[Neu5Ac]1[Fuc]1	17
	3 +	F	1268.8710	ANKTPIRMR + [HexNAc]7[Hex]8	25
8	3 +	A	1272.8513	DGGNNNTNETEIFRPGGGDMR + [HexNAc]3[Hex]5[Fuc]1	91
	3 +	B	1272.8834	VDRNDAGNFTCRATNSVGSK + [HexNAc]5[Hex]3[Fuc]1	17
	3 +	C	1272.8393	VDRNDAGNFTCR + [HexNAc]5[Hex]4[Neu5Ac]2[Fuc]1	11
	3 +	D	1272.9008	KCSKTSFMVENLTGAIWYFR + [HexNAc]2[Hex]6	35
9	3 +	A	1286.5269	DGGNNNTNETEIFRPGGGDMR + [HexNAc]4[Hex]4[Fuc]1	54
	3 +	B	1286.5201	QNLTVKDVTK + [HexNAc]7[Hex]8[Neu5Ac]1[Fuc]1	13
	3 +	C	1286.5467	INGSEPLQVSWYK + [HexNAc]6[Hex]6[Fuc]1	24
	3 +	D	1286.5430	MSDAGKYTVVAGGNVSTAK + [HexNAc]3[Hex]5[Neu5Ac]2	0
10	2 +	A	1298.5581	NCSFNITTEIR + [HexNAc]3[Hex]3[Fuc]1	84
	2 +	B	1298.5734	YTLTLENSSGTK + [HexNAc]4[Hex]2[Fuc]1	28
	2 +	C	1298.5453	YTCQAKNESGVER + [HexNAc]2[Hex]4	25
	2 +	D	1298.5733	YILTVENSSGSK + [HexNAc]4[Hex]3	29
11	3 +	A	1318.5250	DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]8	100
	3 +	B	1318.5804	CSKTSFKVENLTEGAIYYFR + [HexNAc]2[Hex]7	80

	3 +	C	1318.5888	EKNSILWVKLNK + [HexNAc]6[Hex]6[Neu5Ac]1	41
	3 +	D	1318.5902	VQIEKGVNYTQLSIDNCDRNDAGK + [HexNAc]2[Hex]5	65
	3 +	E	1318.5457	KAYANVSSKCSK + [HexNAc]6[Hex]5[Neu5Ac]2	47
12	2 +	A	1327.0688	NCSFNITTEIR + [HexNAc]4[Hex]3	40
	2 +	B	1327.0412	FTNITGEK + [HexNAc]3[Hex]7	17
	2 +	C	1327.0793	LNWTKPEHDGGAK + [HexNAc]2[Hex]4[Fuc]1	0
	2 +	D	1327.1003	RGRQNLTVK + [HexNAc]3[Hex]6	17
13	2 +	A	1335.0606	EANTTLFCASDAK + [HexNAc]3[Hex]3[Fuc]1	96
	2 +	B	1335.0767	LNWTKPEHDGGAK + [HexNAc]2[Hex]5	31
	2 +	C	1335.0257	ANK + [HexNAc]6[Hex]6[Fuc]1	17
	2 +	D	1335.0284	CNITTEK + [HexNAc]2[Hex]8	33
	2 +	E	1335.0684	ANVTVEAR + [HexNAc]5[Hex]4[Fuc]1	26
14	2 +	A	1343.0580	EANTTLFCASDAK + [HexNAc]3[Hex]4	89
	2 +	B	1343.0658	ANVTVEAR + [HexNAc]5[Hex]5	39
	2 +	C	1343.0781	NNVTLK + [HexNAc]6[Hex]3[Fuc]2	0
	2 +	D	1343.0765	YQSNATLVCK + [HexNAc]5[Hex]3	22
15	2 +	A	1367.0686	NCSFNITTEIR + [HexNAc]2[Hex]6	81
	2 +	B	1367.0867	YTFYAGENITSGK + [HexNAc]4[Hex]2[Fuc]1	0
	2 +	C	1367.0839	YTLTLENSSGTK + [HexNAc]3[Hex]5	31
	2 +	D	1367.0764	DGRQNLTVK + [HexNAc]2[Hex]8	31
16	3 +	A	1372.5426	DGGNNNTNETEIFRPGGGDMR + [HexNAc]2[Hex]9	97
	3 +	B	1372.5980	CSKTSFKVENLTEGAIYYFR + [HexNAc]2[Hex]8	81
	3 +	C	1372.6064	EKNSILWVKLNK + [HexNAc]6[Hex]7[Neu5Ac]1	24
	3 +	D	1372.6078	VQIEKGVNYTQLSIDNCDRNDAGK + [HexNAc]2[Hex]6	62
	3 +	E	1372.5896	ESGTTAWQLVNSSVKR + [HexNAc]6[Hex]7	61
17	2 +	A	1387.5819	NCSFNITTEIR + [HexNAc]3[Hex]5	96
	2 +	B	1387.6181	DNGSPILGYWLEK + [HexNAc]4[Hex]2[Fuc]1	0
	2 +	C	1387.5545	FTNITGEK + [HexNAc]2[Hex]9	11
	2 +	D	1387.6022	NGTEILKSK + [HexNAc]4[Hex]6	48
18	2 +	A	1395.6179	SNITGLLLTR + [HexNAc]2[Hex]8	79
	2 +	B	1395.6060	NGINVTSPQR + [HexNAc]6[Hex]3	3
	2 +	C	1395.6156	DNGSPILGYWLEK + [HexNAc]4[Hex]3	4
	2 +	D	1395.5633	ANDTLVR + [HexNAc]3[Hex]5[Neu5Ac]2	6
19	2 +	A	1400.0978	NCSFNITTEIR + [HexNAc]4[Hex]3[Fuc]1	96
	2 +	B	1400.0996	NNVTLK + [HexNAc]8[Hex]3	33
	2 +	C	1400.1051	TKANVTVEAR + [HexNAc]3[Hex]5[Neu5Ac]1	0
	2 +	D	1400.1312	YILKLENSSGSK + [HexNAc]4 + [Hex]4	59
	2 +	E	1400.1293	RGRQNLTVK + [HexNAc]3[Hex]6[Fuc]1	8
20	2 +	A	1416.0820	EANTTLFCASDAK + [HexNAc]3[Hex]4[Fuc]1	80
	2 +	B	1416.1031	LNWTKPEHDGGAK + [HexNAc]2[Hex]6	57
	2 +	C	1416.0948	ANVTVEAR + [HexNAc]5[Hex]5[Fuc]1	23
	2 +	D	1416.0946	VNVSSK + [HexNAc]8[Hex]3	28
21	3 +	A	1451.2518	DGGNNNTNETEIFRPGGGDMR + [HexNAc]5[Hex]4[Neu5Ac]1[Fuc]1	35
	3 +	B	1451.2582	EVNSTHWSRVNK + [HexNAc]5[Hex]8[Neu5Ac]1[Fuc]2	17
	3 +	C	1451.2824	NSLLWKRANK + [HexNAc]7[Hex]6[Neu5Ac]2[Fuc]1	13
	3 +	D	1451.2679	MSDAGKYTVVAGGNVSTAK + [HexNAc]4[Hex]5[Neu5Ac]3	17

	3 +	E	1451.2451	LVINR + [HexNAc]7[Hex]8[Neu5Ac]3[Fuc]1	7
	3 +	F	1451.2328	NNVTLK + [HexNAc]6[Hex]7[Neu5Ac]4[Fuc]1	6
22	2 +	A	1436.6003	EANTTLFCASDAK + [HexNAc]4[Hex]3[Fuc]1	81
	2 +	B	1436.5998	YCVVVENSTGSR + [HexNAc]5[Hex]3	40
	2 +	C	1436.6186	RANHTPESCPEPKYK + [HexNAc]2[Hex]4	8
	2 +	D	1436.6163	LNWTKPEHDGGAK + [HexNAc]3[Hex]5	17
23	2 +	A	1476.6443	SNITGLLLTR + [HexNAc]2[Hex]9	82
	2 +	B	1476.6240	YTFYAGENITSGK + [HexNAc]5[Hex]3	0
	2 +	C	1476.6079	ANVTVEAR + [HexNAc]4[Hex]7[Fuc]1	0
	2 +	D	1476.6321	NGINVTSPQR + [HexNAc]6[Hex]4	3
	2 +	E	1476.6123	YQSNATLVCK <sup>4</sup> + [HexNAc]5[Hex]5	17
24	2 +	A	1481.1242	NCSFNITTEIR + [HexNAc]4[Hex]4[Fuc]1	80
	2 +	B	1481.1576	YILKLENSSGSK + [HexNAc]4[Hex]5	52
	2 +	C	1481.1687	QNATVQGLIQGK + [HexNAc]6[Hex]3	39
	2 +	D	1481.1392	YTVVAGGNVSTAK + [HexNAc]3[Hex]4[Neu5Ac]1[Fuc]1	11
	2 +	E	1481.1443	NGTEILKSK + [HexNAc]5[Hex]5[Fuc]1	32
	2 +	F	1481.1423	VENLTEGAIYYFR + [HexNAc]3[Hex]3[Neu5Ac]1	0
	2 +	G	1481.1395	YTLTLENSSGTK + [HexNAc]5[Hex]3[Fuc]1	34
	2 +	H	1481.1394	YILTVENSSGSK + [HexNAc]5[Hex]4	44
	2 +	I	1481.1315	TKANVTVEAR + [HexNAc]3[Hex]6[Neu5Ac]1	0
	2 +	J	1481.1889	INETLELLSESPVYSTK + [HexNAc]2[Hex]3[Fuc]1	68
25	2 +	A	1517.6265	EANTTLFCASDAK + [HexNAc]4[Hex]4[Fuc]1	72
	2 +	B	1517.6144	ANHTPESCPETK + [HexNAc]5[Hex]4	29
	2 +	C	1517.6162	ANDTLVR + [HexNAc]5[Hex]4[Neu5Ac]2	6
	2 +	D	1517.6388	YQSNATLVCK + [HexNAc]6[Hex]3[Fuc]1	40
26	2 +	A	1568.1958	LINCNTSAITQACPK + [HexNAc]4[Hex]3[Fuc]1	100
	2 +	B	1568.1583	NASGSAKAEIK + [HexNAc]4[Hex]5[Neu5Ac]1[Fuc]1	11
	2 +	C	1568.2197	RESGTTAWQLVNSSVKR + [HexNAc]2[Hex]5	15
	2 +	D	1568.1898	LENSGSKSAFVTVK + [HexNAc]3[Hex]6	17
27	2 +	A	1583.1647	LDVVPIDNNSNYR + [HexNAc]2[Hex]5	100
	2 +	B	1583.1998	VNRLNVTLK + [HexNAc]4[Hex]8	32
	2 +	C	1583.1425	VETNCNLSVEK + [HexNAc]3[Hex]6[Neu5Ac]1	0
	2 +	D	1583.1930	AYATITNCTKTTFR + [HexNAc]3[Hex]4[Fuc]1	0
28	2 +	A	1603.1722	NCSFNITTEIR + [HexNAc]6[Hex]3[Fuc]1	78
	2 +	B	1603.1885	TCILEILNSTK + [HexNAc]4[Hex]5[Neu5Ac]1	7
	2 +	C	1603.1644	YTCQAKNESGVER + [HexNAc]5[Hex]4	46
	2 +	D	1603.1973	QNLTVKDVTK + [HexNAc]4[Hex]5[Neu5Ac]1[Fuc]1	21
29	2 +	A	1610.1478	NCSFNITTEIR + [HexNAc]2[Hex]9	89
	2 +	B	1610.1877	VFAENETGLSRPR + [HexNAc]3[Hex]7	25
	2 +	C	1610.1661	YTFYAGENITSGK + [HexNAc]4[Hex]5[Fuc]1	11
	2 +	D	1610.1137	NASGTK + [HexNAc]4[Hex]5[Neu5Ac]3[Fuc]1	13
30	2 +	A	1669.7355	LINCNTSAITQACPK + [HexNAc]5[Hex]3[Fuc]1	95
	2 +	B	1669.6986	NDAGKYTLTVENSSGSK + [HexNAc]2[Hex]7	5
	2 +	C	1669.7664	DTGEYTLKLVNVTGTTSETIK + [HexNAc]2[Hex]3[Fuc]1	42
	2 +	D	1669.6873	WVRHMK + [HexNAc]6[Hex]7[Fuc]1	24
31	2 +	A	1697.2203	LDVVPIDNNSNYR + [HexNAc]4[Hex]3[Fuc]1	95
	2 +	B	1697.2139	TNKTINHDTQFK + [HexNAc]4[Hex]7	52

	2 +	C	1697.2194	LINCNTSAITQACPK + [HexNAc]2[Hex]8	4
	2 +	D	1697.1902	WVRHMK + [HexNAc]5[Hex]5[Neu5Ac]2[Fuc]1	8
	2 +	E	1697.2506	YTLTVKNASGTK + [HexNAc]8[Hex]3	33
	2 +	F	1697.2122	VDQHEWTKCNTTPTK + [HexNAc]3[Hex]4[Neu5Ac]1	0
	2 +	G	1697.1594	AYACITDNCTK <sup>4</sup> + [HexNAc]6[Hex]6	68
32	2 +	A	1728.2116	NCSFNITTEIR + [HexNAc]5[Hex]4[Neu5Ac]1[Fuc]1	70
	2 +	B	1728.1707	LNK + [HexNAc]6[Hex]7[Neu5Ac]2[Fuc]1	33
	2 +	C	1728.1988	YTCQAKNESGVER + [HexNAc]4[Hex]5[Neu5Ac]1	58
	2 +	D	1728.2150	QSDAGEYTFVAGRNR + [HexNAc]4[Hex]6	33
33	2 +	A	1745.2175	LDVVPIDNNNNSSNYR + [HexNAc]2[Hex]7	92
	2 +	B	1745.2268	CDPPVISNITK + [HexNAc]5[Hex]4[Neu5Ac]2	3
	2 +	C	1745.2541	VNTSPISGREYR + [HexNAc]8[Hex]3	4
	2 +	D	1745.2458	AYATITNNCTKTTFR + [HexNAc]3[Hex]6[Fuc]1	0
34	2 +	A	1826.2439	LDVVPIDNNNNSSNYR + [HexNAc]2[Hex]8	100
	2 +	B	1826.2596	GVNYTQLSIDNCDR + [HexNAc]6[Hex]3[Fuc]2	0
	2 +	C	1826.2532	CDPPVISNITK + [HexNAc]5[Hex]5[Neu5Ac]2	3
	2 +	D	1826.2753	DSGYYSLSAENSSGTDQKIK + [HexNAc]3[Hex]3[Neu5Ac]1	0
	2 +	E	1826.1873	SVDNGHSGRYTCQAKNESGVER + [HexNAc]2[Hex]4[Fuc]1	0

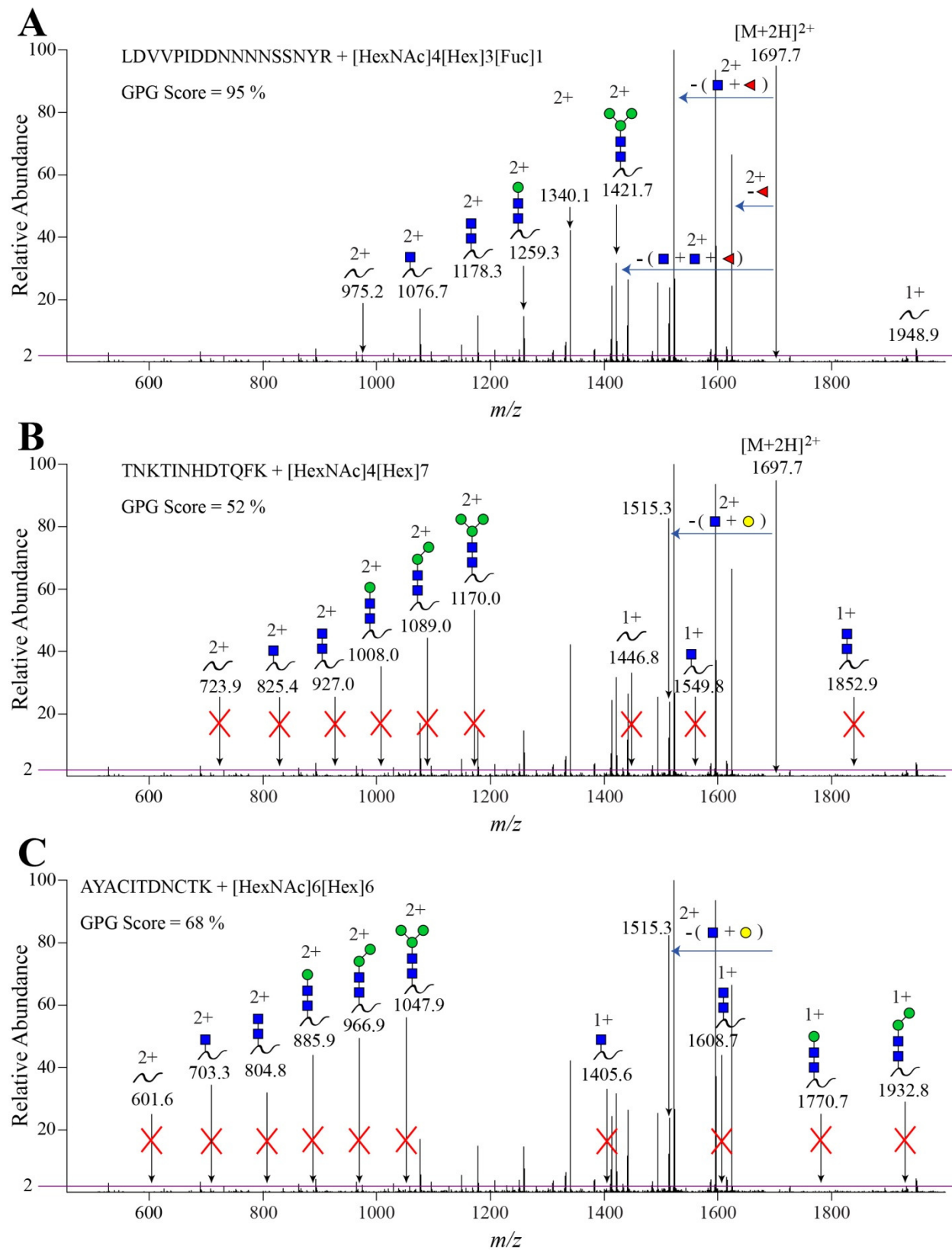
<sup>1</sup> For each test, candidate A is the actual glycopeptide composition that corresponds to the MS/MS data being scored.

<sup>2</sup> All candidate compositions have an *m/z* value, calculated *in silico*, within 50 ppm error of the monoisotopic mass present in the experimental MS<sup>1</sup> data.

<sup>3</sup> 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<sub>1</sub> 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.

<sup>4</sup> 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 gp140 CFI glycopeptide spectrum, scored in the validation data set. An example of a scored fucosylated complex type structure from CON-S gp140 CFI is shown in Figure 3.



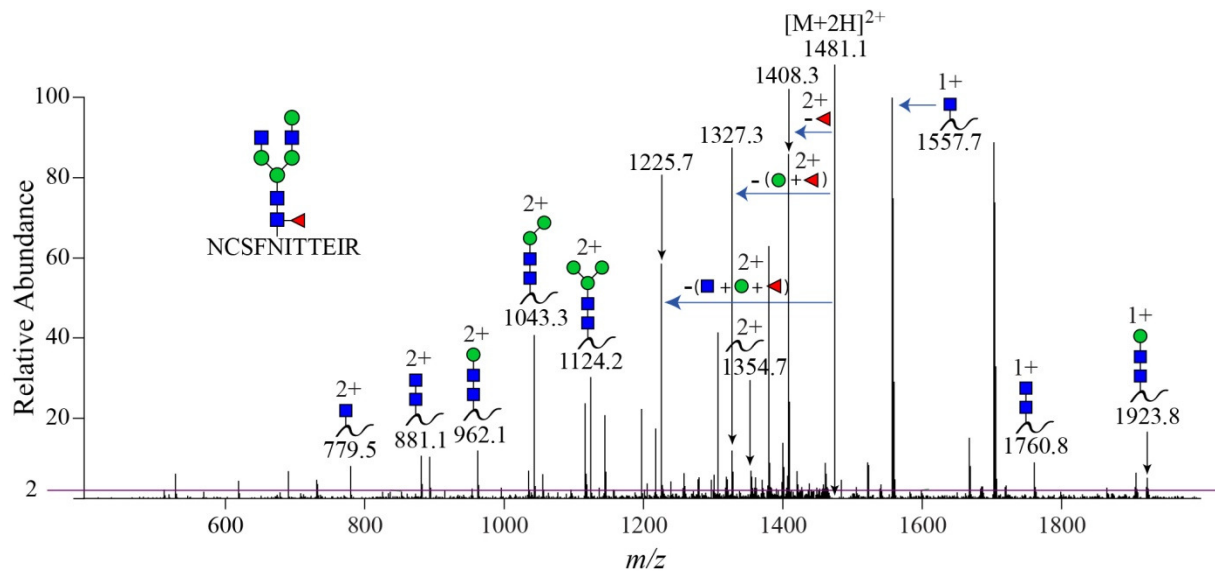
**Figure 3.** MS<sup>2</sup> data from the validation set. (A) GPG evaluation of the correct candidate

**Figure 3.** MS<sup>2</sup> 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<sup>2</sup> 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, INETLELLSESPVYSTK + [HexNAc]2[Hex]3[Fuc]1 was assigned a GPG score of 68 %.

**Table 4.** GPG Results for Candidate Compositions Tested Against the MS<sup>2</sup> Data in Figure 4.

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

<sup>1</sup> 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.

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

<sup>3</sup> 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.

GlycoPepGrader @ glycopro.chem.ku.edu a project of the Heather Desaire Research Group				
Home    Initialize Session    Enter Data    Compute <b>Show Results</b> End/Restart Session				
Peptide	Peptide Score	Glycan	Glycan Score	Candidate Score
SNITGLLLTR	15 / 22 = 0.681818181818182	[Hex]8[HexNAc]2	16 / 16 = 1	0.7868181818181818
NGINVTPSQR	1 / 22 = 0.0454545454545455	[Hex]3[HexNAc]6	0 / 6 = 0	0.0304545454545455
DNGSPILGYWLEK	1 / 16 = 0.0625	[Hex]3[HexNAc]4	0 / 6 = 0	0.041875
ANDTLVR	2 / 21 = 0.0952380952380952	[Hex]5[HexNAc]3[Neu5Ac]2	0 / 8 = 0	0.063809523809524

**Reference:**  
Forthcoming. In the interim, please direct questions to glycopro SWIRL ku PINPOINT edu

**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 SNITGLLLTR + [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

- (1) Apweiler, R.; Hermjakob, H.; Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta-Gen. Subj.* **1999**, *1473*, 4-8.
- (2) Wuhrer, M.; Deelder, A. M.; Hokke, C. H. Protein glycosylation analysis by liquid chromatography-mass spectrometry. *J. Chromatogr. B.* **2005**, *825*, 124-133.
- (3) Morelle, W.; Canis, K.; Chirat, F.; Faid, V.; Michalski, J. C. The use of mass spectrometry for the proteomic analysis of glycosylation. *Proteomics.* **2006**, *6*, 3993-4015.
- (4) Murrell, M. P.; Yarema, K. J.; Levchenko, A. The systems biology of glycosylation. *ChemBiochem.* **2004**, *5*, 1334-1347.
- (5) Helenius, A.; Aebi, M. Intracellular functions of *N*-linked glycans. *Science.* **2001**, *291*, 2364-2369.
- (6) Petrescu, A. J.; Wormald, M. R.; Dwek, R. A. Structural aspects of glycomes with a focus on *N*-glycosylation and glycoprotein folding. *Curr. Opin. Struct. Biol.* **2006**, *16*, 600-607.
- (7) Skropeta, D. The effect of individual *N*-glycans on enzyme activity. *Bioorganic & Medicinal Chemistry.* **2009**, *17*, 2645-2653.
- (8) Budnik, B. A., Lee, R. S.; Steen, J. A. J. Global methods for protein glycosylation analysis by mass spectrometry. *BBA-Protein Proteomics.* **2006**, *1764*, 1870-1880.
- (9) Bertozzi, C. R.; Kiessling, L. L. Chemical glycobiology. *Science.* **2001**, *291*, 2357-2364.
- (10) Zaia, J. Mass spectrometry and the emerging field of glycomics. *Chemistry & Biology.* **2008**, *15*, 881-892.
- (11) Drake P. M.; Cho, W.; Li, B.; Prakobphol, A.; Johansen, E.; Anderson, N. L.; Regnier, F.E.; Gibson, B.W.; Fisher S. J. Sweetening the pot: Adding glycosylation to the biomarker discovery equation. *Clin. Chem.* **2010**, *56*, 223-236.
- (12) Dalpathado, D. S.; Desaire, H. Glycopeptide analysis by mass spectrometry. *Analyst.* **2008**, *133*, 731-738.
- (13) Cooper, C. A.; Gasteiger, E.; Packer, N. H. GlycoMod – A software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics.* **2001**, *1*, 340-349.
- (14) Go, E. P.; Rebecchi, K. R.; Dalpathado, D. S.; Bandu, M. L.; Zhang, Y.; Desaire, H. GlycoPep DB: A tool for glycopeptide analysis using a "smart search". *Anal. Chem.* **2007**, *79*, 1708-1713.

- (15) Desaire, H.; Hua, D. When can glycopeptides be assigned based solely on high-resolution mass spectrometry data? *Int. J. Mass. Spectrom.* **2009**, *287*, 21-26.
- (16) Wang, X.; Emmett, M. R.; Marshall, A. G. Liquid chromatography electrospray ionization Fourier transform ion cyclotron resonance mass spectrometric characterization of *N*-linked glycans and glycopeptides. *Anal. Chem.* **2010**, *82*, 6542-6548.
- (17) Wuhrer, M.; Catalina, M. I.; Deelder, A. M.; Hokke, C. H. Glycoproteomics based on tandem mass spectrometry of glycopeptides. *J. Chromatogr. B.* **2007**, *849*, 115-128.
- (18) Ren, J. M.; Rejtar, T.; Li, L.; Karger, B. L. *N*-glycan structure annotation of glycopeptides using a linearized glycan structure database (GlyDB). *J. Proteome Res.* **2007**, *6*, 3162-3173.
- (19) Joenväärä, S.; Ritamo, I.; Peltoniemi, H.; Renkonen, R. *N*-Glycoproteomics – An automated workflow approach. *Glycobiology.* **2008**, *18*, 339-349.
- (20) Shan, B.; Ma, B.; Zhang, K.; Lajoie, G. Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *J. Bioinform. Comput. Biol.* **2008**, *6*, 77-91.
- (21) Peltoniemi, H.; Joenväärä, S.; Renkonen, R. *De novo* glycan structure search with the CID MS/MS spectra of native *N*-glycopeptides. *Glycobiology.* **2009**, *19*, 707-714.
- (22) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H.R.; Dell, A. Automated *N*-glycopeptide identification using a combination of single- and tandem-MS. *J. Proteome Res.* **2007**, *6*, 3995-4005.
- (23) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. GlycoWorkbench: A tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.* **2008**, *7*, 1650-1659.
- (24) Maass, K.; Ranzinger, R.; Geyer, H.; von der Lieth, C. W.; Geyer, R. "Glyco-peakfinder" – *De novo* composition analysis of glycoconjugates. *Proteomics.* **2007**, *7*, 4435-4444.
- (25) Irungu, J.; Go, E. P.; Dalpathado, D. S.; Desaire, H. Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID. *Anal. Chem.* **2007**, *79*, 3065-3074.
- (26) Ozohanics, O.; Krenyacz, J.; Ludányi, K.; Pollreisz F.; Vékey, K.; Drahos, L. GlycoMiner: A new software tool to elucidate glycopeptide composition. *Rapid Commun. Mass. Spectrom.* **2008**, *22*, 3245-3254.
- (27) Go, E. P.; Irungu, J.; Zhang, Y.; Dalpathado, D. S.; Liao, H. X.; Sutherland, L. L.; Alam, S. M.; Haynes, B. F.; Desaire, H. Glycosylation site-specific analysis of HIV envelope proteins (JR-FL and CON-S) reveals major differences in glycosylation site occupancy, glycoform profiles, and antigenic epitopes' accessibility. *J. Proteome Res.* **2008**, *7*, 1660-1674.
- (28) Gao, F.; Weaver, E. A.; Lu, Z.; Li, Y.; Liao, H. X.; Ma, B.; Alam, S. M.; Scarce, R. M.;

Sutherland, L. L.; Yu, J. S.; Decker, J. M.; Shaw, G. M.; Montefiori, D. C.; Korber, B. T.; Hahn, B. H.; Haynes, B. F. Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group M consensus envelope glycoprotein. *J. Virol.* **2005**, *79*, 1154-1163.

(29) Liao, H. X.; Sutherland, L. L.; Xia, S. M.; Brock, M. E.; Scarce, R. M.; Vanleeuwen, S.; Alam, S. M.; McAdams, M.; Weaver, E. A.; Camacho, Z. T.; Ma, B. J.; Li, Y.; Decker, J. M.; Nabel, G. J.; Montefiori, D. C.; Hahn, B. H.; Korber, B. T.; Gao, F.; Haynes, B. F. A group M consensus envelope glycoprotein induces antibodies that neutralize subsets of subtype B and C HIV-1 primary viruses. *Virology.* **2006**, *353*, 268-282.

(30) Rebecchi, K.R.; Wenke, J. L.; Go, E.P.; Desaire, H. Label-free quantitation: A new glycoproteomics approach. *J. Am. Soc. Mass. Spectrom.* **2009**, *20*, 1048-1059.

(31) Alley, W. R.; Mechref, Y.; Novotny, M. V. Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. *Rapid Commun. Mass. Spectrom.* **2009**, *23*, 161-170.

(32) Satomi, Y.; Shimonishi, Y.; Hase, T.; Takao, T. Site-specific carbohydrate profiling of human transferrin by nano-flow liquid chromatography/electrospray ionization mass spectrometry. *Rapid Commun. Mass. Spectrom.* **2004**, *18*, 2983-2988.

(33) Ritchie, M. A.; Gill, A. C.; Deery, M. J.; Lilley, K. Precursor ion scanning for detection and structural characterization of heterogeneous glycopeptide mixtures. *J. Am. Soc. Mass. Spectrom.* **2002**, *13*, 1065-1077.

## 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.<sup>1, 2, 3, 4, 5</sup> Experimental studies show that in the absence of proper disulfide bond orientation, proper protein folding may not occur.<sup>2, 6</sup> 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.<sup>2, 7, 8, 9</sup> 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.<sup>7, 9, 10, 11, 12, 13</sup>

At present, mass spectrometry is a common analysis route used to identify disulfide bond arrangements in proteins and peptides.<sup>14</sup> Nuclear magnetic resonance (NMR)<sup>15, 16</sup> and crystallography<sup>17, 18, 19</sup> 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,<sup>9, 10, 13, 14, 20, 21</sup> 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,<sup>14</sup> including collision induced dissociation (CID).<sup>21, 22, 23, 24, 25, 26</sup> However, ETD MS/MS impart more extensive fragmentation information for these species.<sup>11, 25</sup>

During ETD MS/MS experiments, C -N bond cleavage is induced through the transfer of an electron from a radical anion, such as fluoranthene, to a protonated peptide.<sup>11, 27</sup> The mechanism of peptide backbone cleavage is analogous to the way in which fragmentation in electron capture dissociation (ECD) occurs.<sup>11, 28, 29</sup> When the backbone is fragmented, it dissociates into c- and z-type product ions.<sup>11, 27, 30, 31</sup> Like the b- and y-type ions generated

during collision induced dissociation (CID),<sup>11, 27, 30</sup> 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),<sup>11, 27, 30</sup> it is shown to *preferentially* cleave between peptides containing a disulfide bond.<sup>10, 12, 13, 30</sup> 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,<sup>11, 27, 30</sup> it does not typically fragment the covalent disulfide bond.<sup>10, 11</sup>

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.<sup>32, 33</sup> 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.<sup>34, 35, 36</sup> Two of these programs are accessible to only select users. These include the commercially available Charger, developed by Sadygov *et al.* and distributed by Thermo Scientific, and Charge Prediction Machine (CPM) by Carvalho *et al.*, which is available to academic users to run on Linux through the Mono Project.<sup>35, 36</sup> The other program, developed by Sharma *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.<sup>34</sup> 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.<sup>12</sup> 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<sub>3</sub> OH) and HPLC grade acetonitrile (CH<sub>3</sub> CN) were purchased from Fisher Scientific (Fairlawn, NJ). Ammonium bicarbonate (NH<sub>4</sub>HCO<sub>3</sub>) 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<sub>4</sub>HCO<sub>3</sub> (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](http://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.

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

<sup>1</sup> Lysozyme 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.

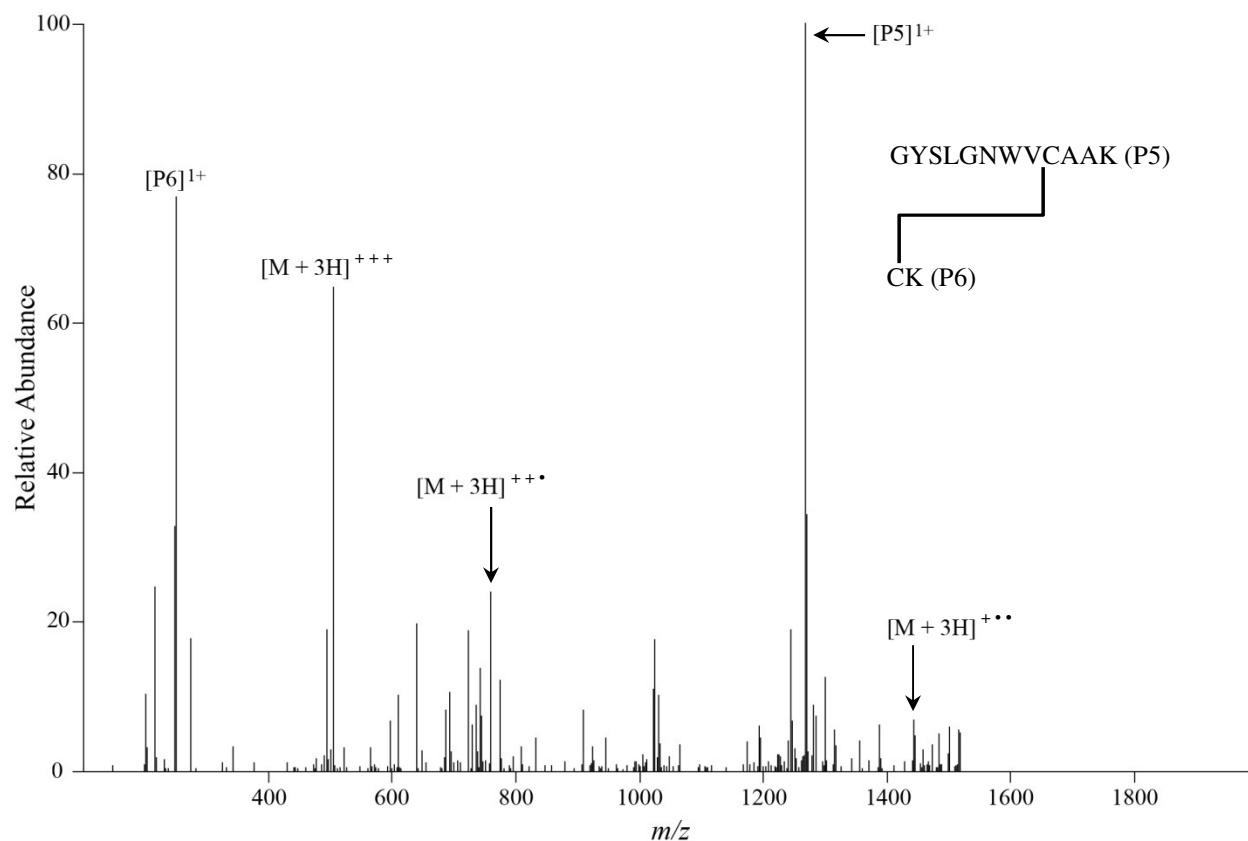
Currently, there are a limited number of programs that exist to determine charge state from low resolution ETD MS/MS data, and only one of these is publicly accessible.<sup>34</sup> This program, published by MacCoss and co-workers takes advantage of a characteristic feature found in ETD spectra, the presence of intense peaks corresponding to charge reduced precursor species.<sup>34</sup>

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.<sup>12, 30</sup>

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.



**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).<sup>37</sup> This is shown below by Equation 1.



**Equation 1.** Charge State ( $z$ ) of Peak 1 = Peak 1 / (Peak 2 – Peak 1) + 1

As charge reduced precursor ions are readily detectable within the ETD MS/MS data of disulfide-bonded peptides,<sup>29</sup> 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 straight forward 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.**  $F_x$  (Column C) = Column A/(Column B – 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.

	A	B	C
1	<b><u>Peak 1</u></b>	<b><u>Peak 2</u></b>	<b><u>Charge Value</u></b>
2			
3	818.03	1090.79	2.999
4	818.03	1091.54	2.991
5	818.03	1620.30	1.020
6	818.03	1622.43	1.017
7	818.03	1623.41	1.016
8	818.03	1627.38	1.011
9	818.03	1628.35	1.010
10	818.03	1629.19	1.008
11	818.03	1635.95	1.000
12	818.03	1637.66	0.998
13			
14			

**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.

**Equation 3.**  $F_x$  (Column B) = ABS(Column A-ROUND(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).

**Equation 4.**  $F_x$  (Column C) = IF(ABS(Column B)>0.01,0,ROUND(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.

	A	B	C
1			
2	<b><u>Charge Value</u></b>	<b><u>Rounding Distance</u></b>	<b><u>Charge State</u></b>
3			
4	2.999	0.0009	3
5	0.999	0.0010	1
6	0.998	0.0020	1
7	1.008	0.0085	1
8	2.991	0.0091	3
9	1.010	0.0095	1
10	1.011	0.0107	0
11	1.016	0.0157	0
12	1.017	0.0169	0
13	1.020	0.0196	0
14			
15			

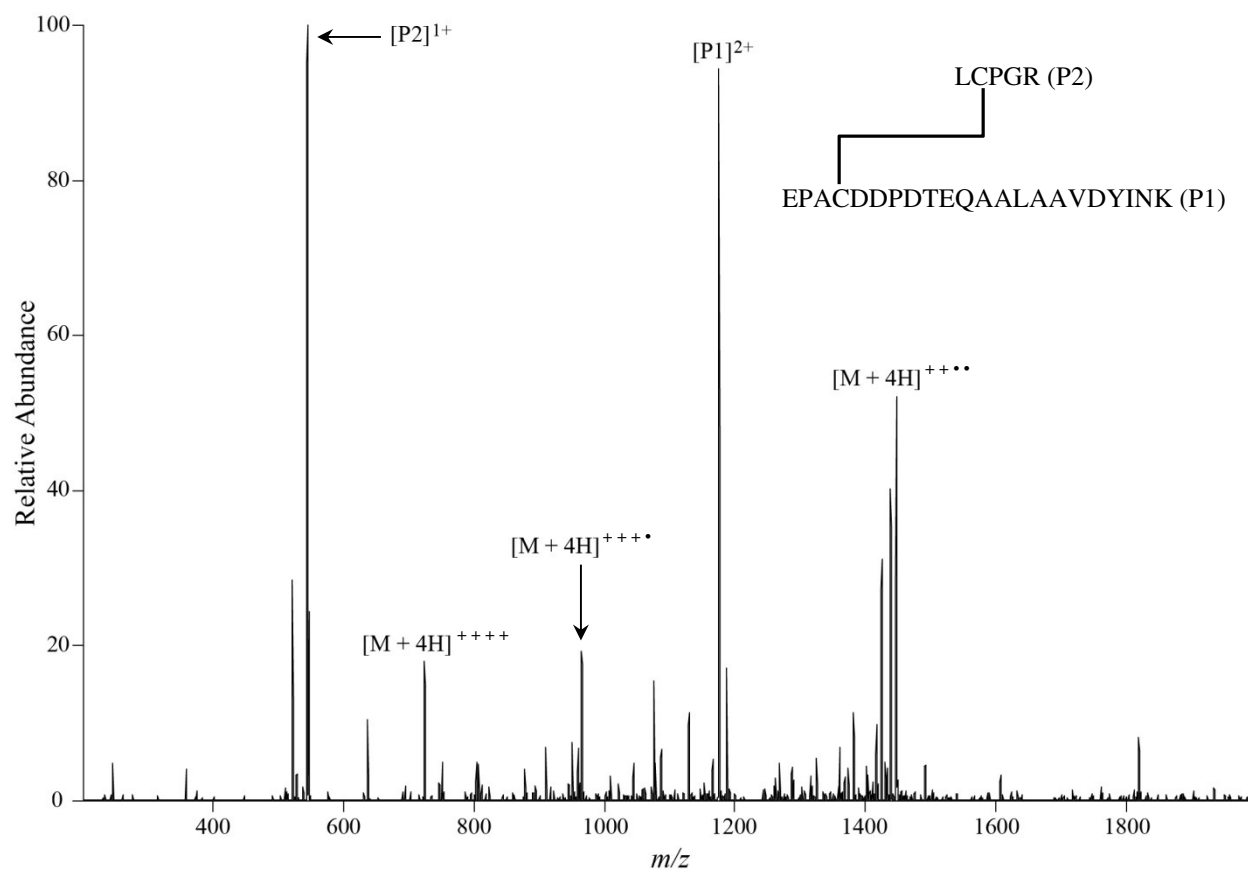
**Figure 3.** Screen shot depicting computational tool 2. The charge values shown in Column A are the output from computational tool 2. For the MS/MS data shown here, a charge state 3 + for the charge reduced precursor species in the precursor – 1 charge state, the disulfide bonded precursor is a 4 + 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.<sup>10, 12, 13,</sup>

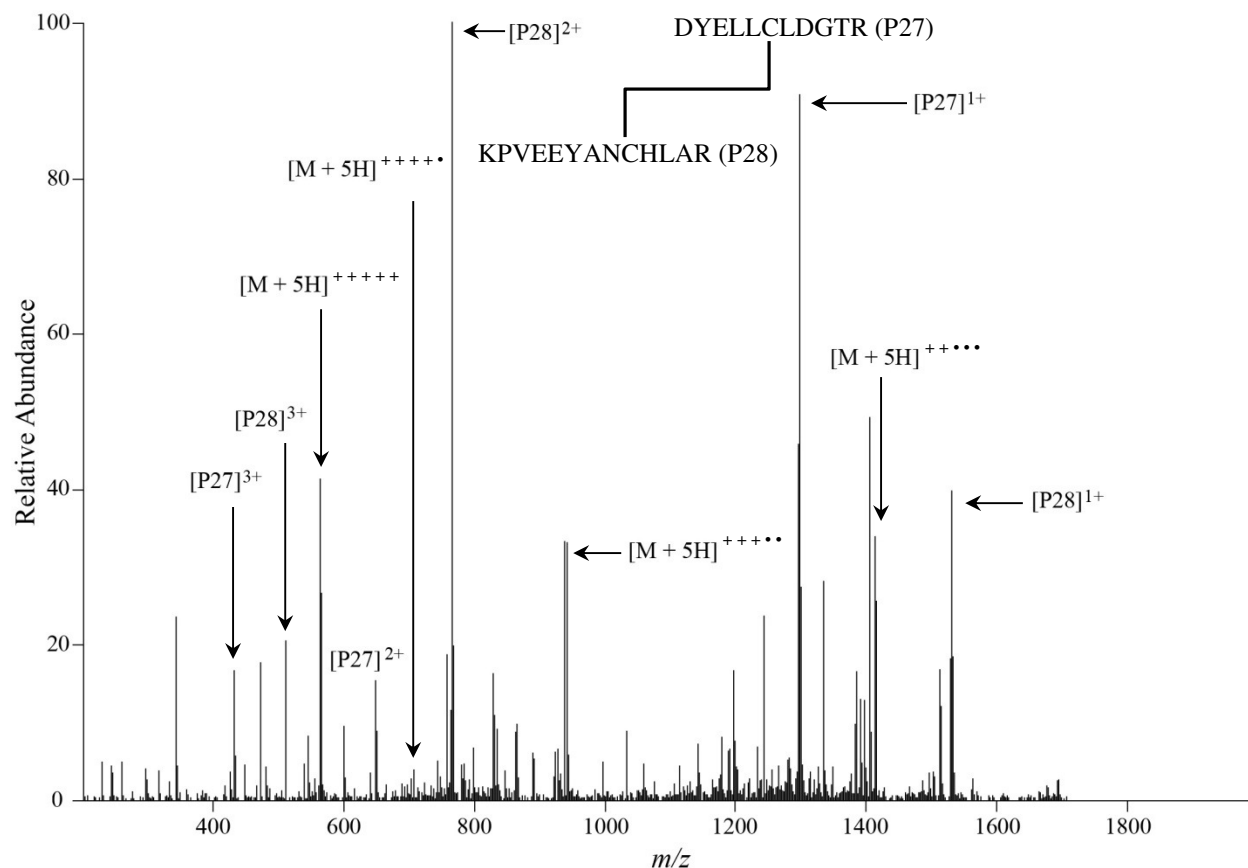
30



**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.

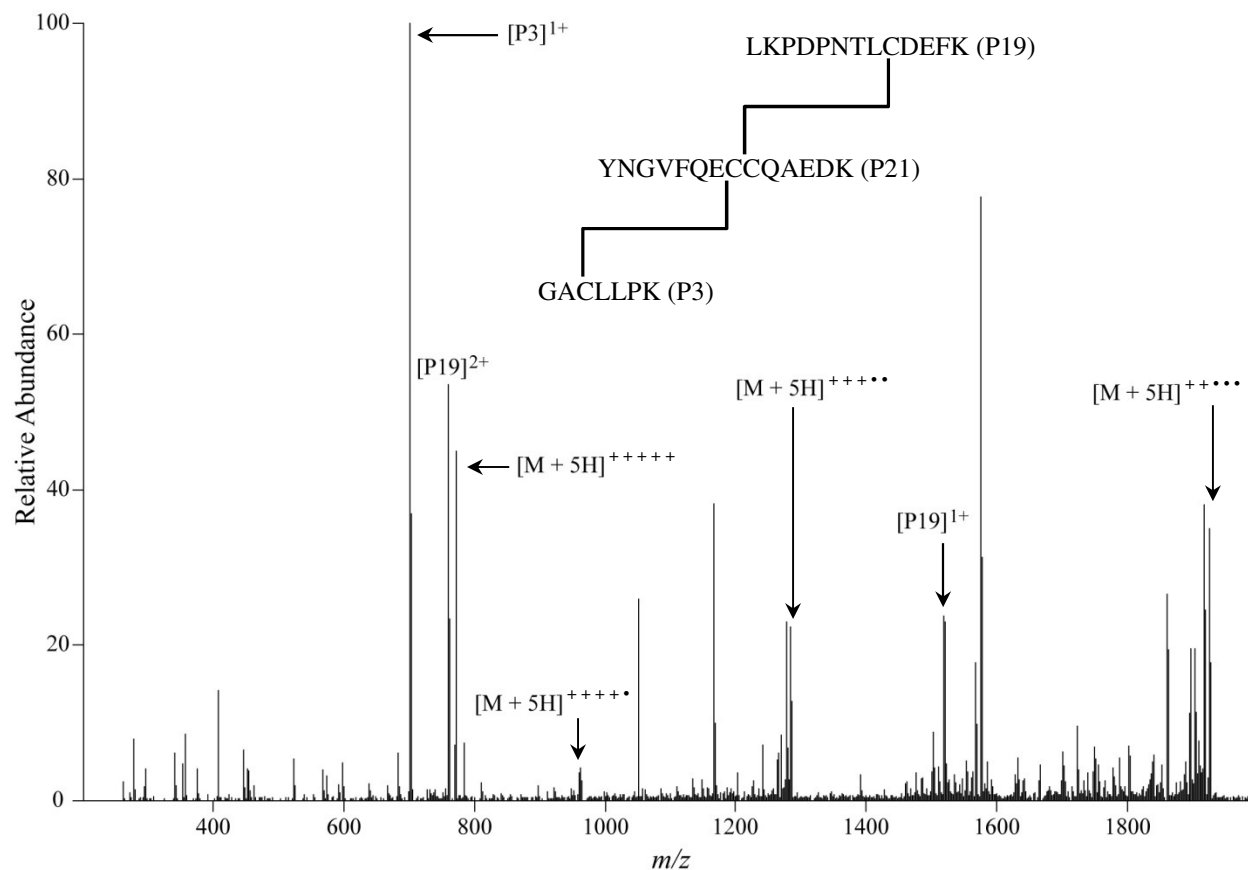
One of the initial difficulties in designing a method for precursor charge state

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.<sup>34</sup> 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.<sup>34</sup> 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.



**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.

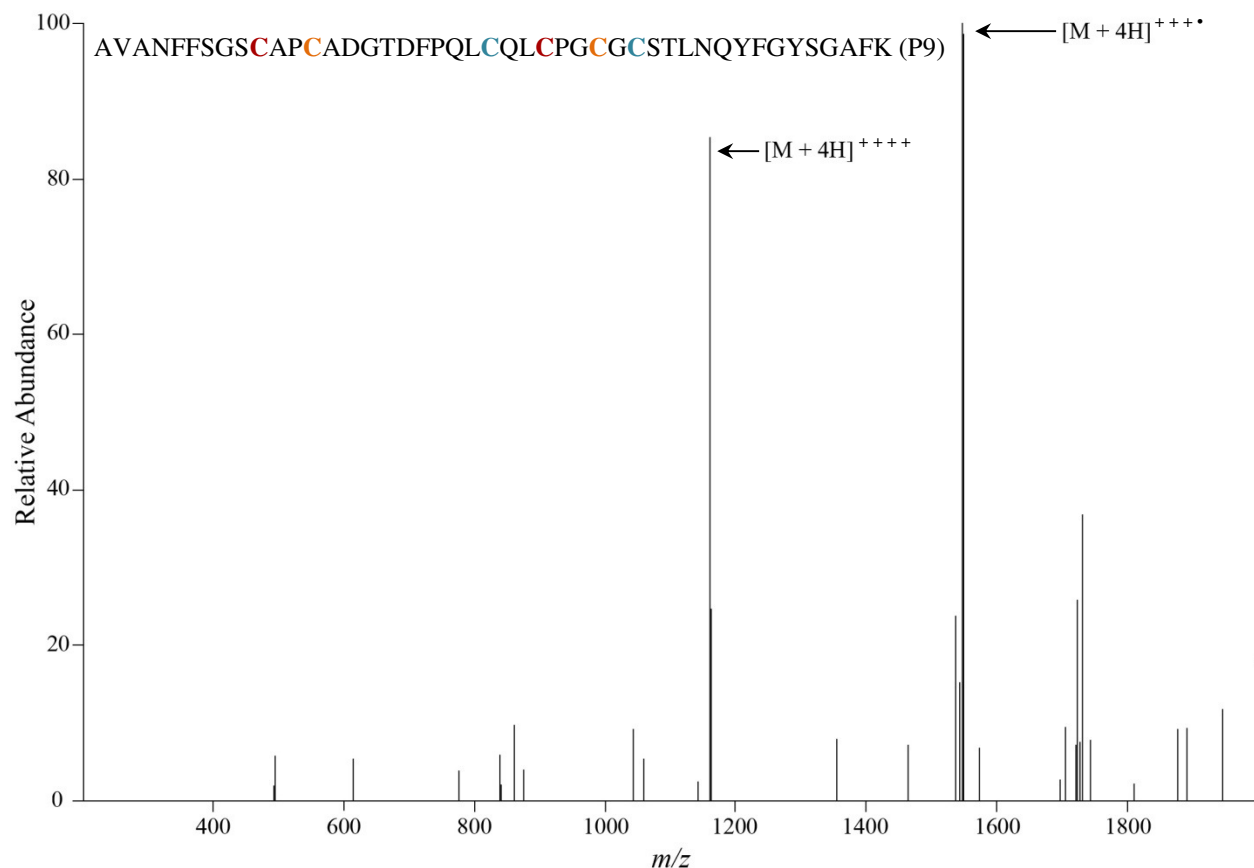
Often, more than two individual chains are joined to form a precursor containing native disulfide bonding. Therefore, it was important to ensure that our charge state determination tools were applicable to these disulfide-bonded peptides as well. The ETD spectrum collected on a precursor from BSA that 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 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.

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.<sup>12</sup> These include variations seen in both the amount, and type, of product ions formed.<sup>12</sup> 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 7.** ETD MS/MS data at  $m/z$  1161.26 from a 4 + charge state transferrin precursor with three intrachain disulfide bonds. The cysteine residues forming each disulfide bond are shown in color within the peptide sequence.

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.

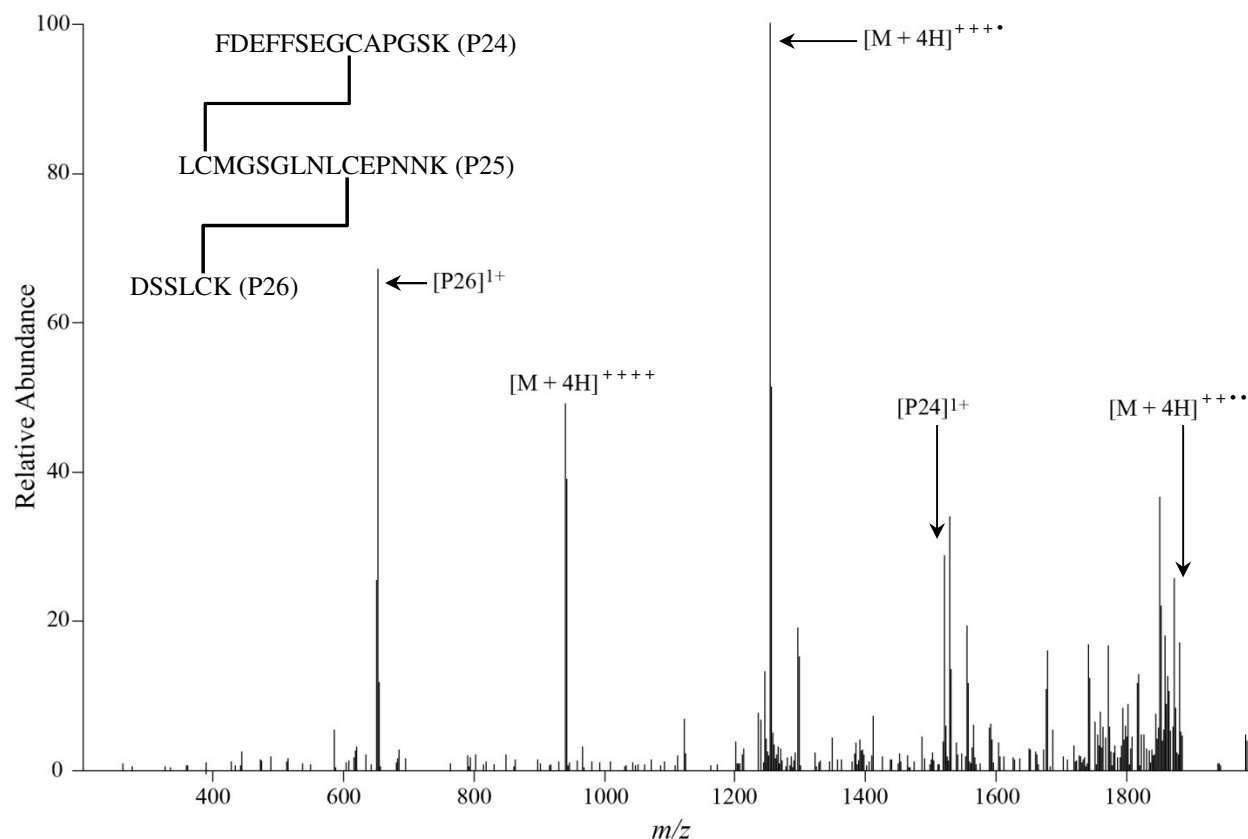
**4.3.4 Number of Charge State Assignments Returned.** One of the limitations to current programs designed for charge state determination of peptide ETD MS/MS data is the inability to assign a single charge state to a precursor ion. The number of charge states returned to a user per spectrum tested is 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.<sup>34</sup> 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.<sup>34</sup> This means that the meaning SVM classifier will aim to generate 1.53 charge state predictions for each spectral search performed.<sup>34</sup> 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.<sup>34</sup> 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. On 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 %.<sup>34</sup> 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 %.<sup>34</sup>

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.



**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

- (1) Fass, D. Disulfide bonding in protein biophysics. In *Annual Reviews of Biophysics, Vol. 41*, Rees, D. C., Ed. 2012; Vol. 41, pp 63-79.
- (2) Trivedi, M. V.; Laurence, J. S.; Siahaan, T. J. The role of thiols and disulfides on protein stability. *Curr. Protein Pept. Sci.* **2009**, *10*, 614-625.
- (3) Kang, T. S.; Kini, R. M. Structural determinants of protein folding. *Cell. Mol. Life Sci.* **2009**, *66*, 2341-2361.
- (4) Wong, J. W. H.; Ho, S. Y. W.; Hogg, P. J. Disulfide bond acquisition through eukaryotic protein evolution. *Mol. Biol. Evol.* **2011**, *28*, 327-334.
- (5) Feige, M. J.; Hendershot, L. M. Disulfide bonds in ER protein folding and homeostasis. *Curr. Opin. Cell Biol.* **2011**, *23*, 167-175.
- (6) Wedemeyer, W. J.; Welker, E.; Narayan, M.; Scheraga, H. A. Disulfide bonds and protein folding. *Biochemistry.* **2000**, *39*, 4207-4216.
- (7) Chen, G.; Warrack, B. M.; Goodenough, A. K.; Wei, H.; Wang-Iverson, D. B.; Tymiak, A. A. Characterization of protein therapeutics by mass spectrometry: Recent developments and future directions. *Drug Discov. Today.* **2011**, *16*, 58-64.
- (8) Zhang, L.; Chou, C. P.; Moo-Young, M. Disulfide bond formation and its impact on the biological activity and stability of recombinant therapeutic proteins produced by *Escherichia coli* expression system. *Biotechnol. Adv.* **2011**, *29*, 923-929.
- (9) Wu, S. L.; Jiang, H.; Lu, Q.; Dai, S.; Hancock, W. S.; Karger, B. L. Mass spectrometric determination of disulfide linkages in recombinant therapeutic proteins using online LC-MS with electron-transfer dissociation. *Anal. Chem.* **2009**, *81*, 112-122.
- (10) Wang, Y.; Lu, Q.; Wu, S. L.; Karger, B. L.; Hancock, W. S. Characterization and comparison of disulfide linkages and scrambling patterns in therapeutic monoclonal antibodies: Using LC-MS with electron transfer dissociation. *Anal. Chem.* **2011**, *83*, 3133-3140.
- (11) Mikesch, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E. P.; Shabanowitz, J.; Hunt, D. F. The utility of ETD mass spectrometry in proteomic analysis. *BBA-Proteins Proteomics.* **2006**, *1764*, 1811-1822.
- (12) Cole, S. R.; Ma, X.; Zhang, X.; Xia, Y. Electron transfer dissociation (ETD) of peptides containing intrachain disulfide bonds. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 310-320.
- (13) Clark, D. F.; Go, E. P.; Desaire, H. Simple approach to assign disulfide connectivity using extracted ion chromatograms of electron transfer dissociation spectra. *Anal. Chem.* **2013**, *85*, 1192-1199.

- (14) Gorman, J. J.; Wallis, T. P.; Pitt, J. J. Protein disulfide bond determination by mass spectrometry. *Mass Spectrom. Rev.* **2002**, *21*, 183-216.
- (15) Sharma, D.; Rajarathnam, K.  $^{13}\text{C}$  NMR chemical shifts can predict disulfide bond formation. *J. Biomol. NMR.* **2000**, *18*, 165-171.
- (16) Mobli, M.; King, G. F. NMR methods for determining disulfide-bond connectivities. *Toxicon.* **2010**, *56*, 849-854.
- (17) McGeehan, J. E.; Bourgeois, D.; Royant, A.; Carpentier, P. Raman-assisted crystallography of biomolecules at the synchrotron: Instrumentation, methods and applications. *BBA-Proteins Proteomics.* **2011**, *1814*, 750-759.
- (18) Qiu, W.; Dong, A.; Pizarro, J. C. Botchkarsev, A.; Min, J.; Wernimont, A. K.; Hills, T.; Hui, R.; Artz, J. D. Crystal structures from the *Plasmodium* peroxiredoxins: New insights into oligomerization and product binding. *BMC Struct. Biol.* **2012**, *12*.
- (19) Hwang, S.; Hilty, C. Folding determinants of disulfide bond forming protein B explored by solution nuclear magnetic resonance spectroscopy. *Proteins.* **2011**, *79*, 1365-1375.
- (20) Yen, T. Y.; Joshi, R. K.; Yan, H.; Seto, N. O. L.; Palcic, M. M.; Macher, B. A. Characterization of cysteine residues and disulfide bonds in proteins by liquid chromatography/electrospray ionization tandem mass spectrometry. *J. Mass Spectrom.* **2000**, *35*, 990-1002.
- (21) Wu, S. L.; Jiang, H.; Hancock, W. S.; Karger, B. L. Identification of the unpaired cysteine status and complete mapping of the 17 disulfides of recombinant tissue plasminogen activator using LC-MS with electron transfer dissociation/collision induced dissociation. *Anal. Chem.* **2010**, *82*, 5296-5303.
- (22) Clark, D. F.; Go, E. P.; Toumi, M. L.; Desaire, H. Collision induced dissociation products of disulfide-bonded peptides: Ions result from the cleavage of more than one bond. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 492-498.
- (23) Chen, J.; Shiyanov, P.; Zhang, L.; Schlager, J. J.; Green-Church, K. B. Top-down characterization of a native highly intralinked protein: Concurrent cleavages of disulfide and protein backbone bonds. *Anal. Chem.* **2010**, *82*, 6079-6089.
- (24) Janecki, D. J.; Nemeth, J. F. Application of MALDI TOF/TOF mass spectrometry and collision-induced dissociation for the identification of disulfide-bonded peptides. *J. Mass Spectrom.* **2011**, *46*, 677-688.
- (25) Mentinova, M.; Hongling, H.; McLuckey, S. A. Dissociation of disulfide-intact somatostatin ions: The roles of ion type and dissociation method. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 2647-2655.

- (26) Mormann, M.; Eble, J.; Schwöppe, C.; Mesters, R. M.; Berdel, W. E.; Peter-Katalini, J.; Pohlentz, G. Fragmentation of intra-peptide and inter-peptide disulfide bonds of proteolytic peptides by nanoESI collision-induced dissociation. *Anal. Bioanal. Chem.* **2008**, *392*, 831-838.
- (27) Molina, H.; Matthiesen, R.; Kandasamy, K.; Pandey, A. Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Anal. Chem.* **2008**, *80*, 4825-4835.
- (28) Zubarev, R. A. Electron-capture dissociation tandem mass spectrometry. *Curr. Opin. Biotechnol.* **2004**, *15*, 12-16.
- (29) Sun, R. X.; Dong, M. Q.; Song, C. Q.; Chi, H.; Yang, B.; Xiu, L. Y.; Tao, L.; Jing, Z. Y.; Liu, C.; Wang, L. H.; Fu, Y.; He, S. M. Improved peptide identification for proteomic analysis based on comprehensive characterization of electron transfer dissociation spectra. *J. Proteome Res.* **2010**, *9*, 6354-6367.
- (30) Wiesner, J.; Premisler, T.; Sickmann, A. Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics.* **2008**, *8*, 4466-4483.
- (31) Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics.* **2007**, *6*, 1942-1951.
- (32) Xu, H.; Zhang, L.; Freitas, M. A. Identification and characterization of disulfide bonds in proteins and peptides from tandem MS data by use of the MassMatrix MS/MS search engine. *J. Proteome Res.* **2008**, *7*, 138-144.
- (33) Xu, H.; Hsu, P. H.; Zhang, L.; Tsai, M. D.; Freitas, M. A. Database search algorithm for identification of intact cross-links in proteins and peptides using tandem mass spectrometry. *J. Proteome Res.* **2010**, *9*, 3384-3393.
- (34) Sharma, V.; Eng, J. K.; Feldman, S.; von Haller, P. D.; MacCoss, M. J.; Noble, W. S. Precursor charge state prediction for electron transfer dissociation tandem mass spectra. *J. Proteome Res.* **2010**, *9*, 5438-5444.
- (35) Sadygov, R. G.; Hao, Z.; Huhmer, A. F. R. Charger: Combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. *Anal. Chem.* **2008**, *80*, 376-386.
- (36) Carvalho, P. C.; Cociorva, D.; Wong, C. C. L.; Carvalho, M. D. D.; Barbosa, V. C.; Yates, J. R. Charge prediction machine: Tool for inferring precursor charge states of electron transfer dissociation tandem mass spectra. *Anal. Chem.* **2009**, *81*, 1996-2003.
- (37) Chassigne, H.; Vacchina, V.; Łobiński, R. Elemental speciation analysis in biochemistry by electrospray mass spectrometry. *Trac-Trends Anal. Chem.* **2000**, *19*, 300-313.



## 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 to 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).<sup>1, 2, 3, 4</sup> Although the negative charges add complexity to their analysis, these acidic glycans play major functional roles in biological processes.<sup>4, 5, 6, 7, 8</sup> In addition, sialic acids have proven critical to the development of efficacious and safe glycoprotein therapeutics.<sup>6, 9, 10, 11, 12, 13</sup> 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.<sup>6, 9, 10, 11, 12</sup>

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.<sup>14</sup> 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.<sup>14, 15</sup> 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.<sup>15</sup> 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.<sup>16</sup> 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.<sup>15</sup>

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.<sup>17, 18</sup>

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 proposed 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 (CH<sub>3</sub>OH) and HPLC grade acetonitrile (CH<sub>3</sub>CN) were purchased from Fisher Scientific (Fairlawn, NJ). Ammonium bicarbonate (NH<sub>4</sub>HCO<sub>3</sub>) 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 M .

### **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*

*al.*<sup>19</sup> 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.<sup>15</sup>

**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.*<sup>15, 20</sup> Sample preparation for the CON-S gp140 CFI glycopeptides was also performed by Go and co-workers, as previously described.<sup>20</sup> Detailed information on the experimental procedures and MS analysis is also given in Chapter 3 of this dissertation.

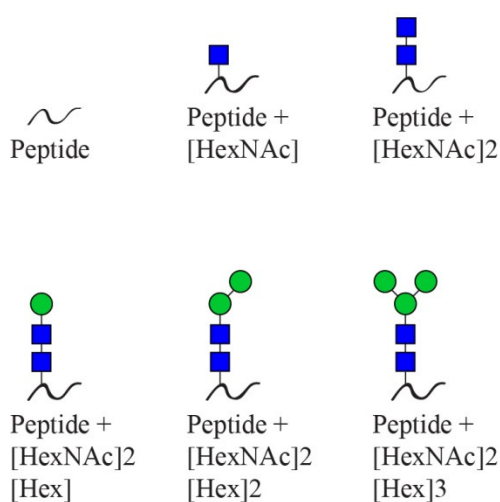
**5.2.4 Manual Data Analysis.** To identify the glycopeptides from these samples in the MS data, a prediction table of theoretical  $m/z$  values corresponding to glycopeptide compositions for RNase B, asialofetuin, and transferrin was prepared. The amino acid sequences from the proteins were obtained from Uniprot ([www.uniprot.org](http://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 carbamidomethylation, 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.

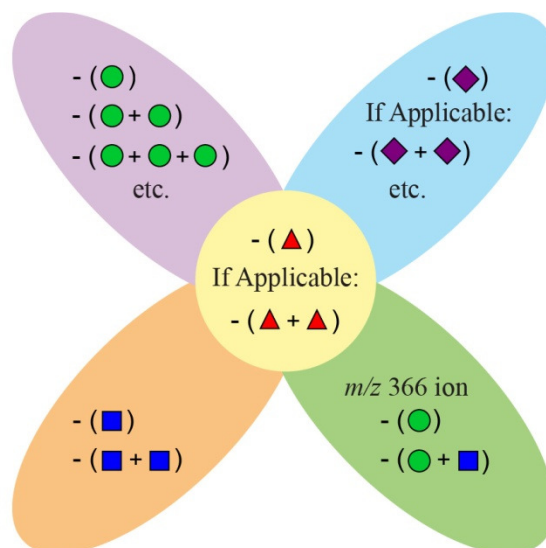
### **5.3 RESULTS AND DISCUSSION**

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.<sup>15</sup>

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

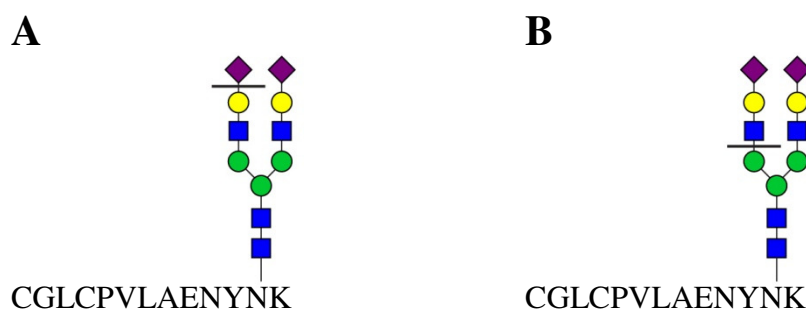


corresponded to the [peptide + HexNAc], [peptide + 2HexNAc], [peptide + 2HexNAc + Hex], and [peptide + 2 HexNAc + 2Hex] ions, but no product ions was detected for the calculated [peptide] or [peptide + 2 HexNAc + 3Hex] ions, the longest number of consecutive matches for the peptide-containing product ions would be 4. Therefore, 4 points would be added to that glycopeptide's *ActualRawPeptideScore*. However, since there were a total of 6 peptide-containing product ions possible, 6 points would be added to that candidate's *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 take place for these species under different charge states.<sup>14, 15, 17, 21</sup> 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.<sup>14</sup> However, glycosidic cleavages may result from two distinct mechanisms, charge-remote pathway or charge-directed pathway.<sup>14, 17</sup> Proton distribution is shown to affect the probability of each pathway, especially in the case of sialylated glycopeptide precursors.<sup>14</sup> For these species, it has recently been shown that precursor charge state determines which of the cleavage pathways will dominate.<sup>14</sup>

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 *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.

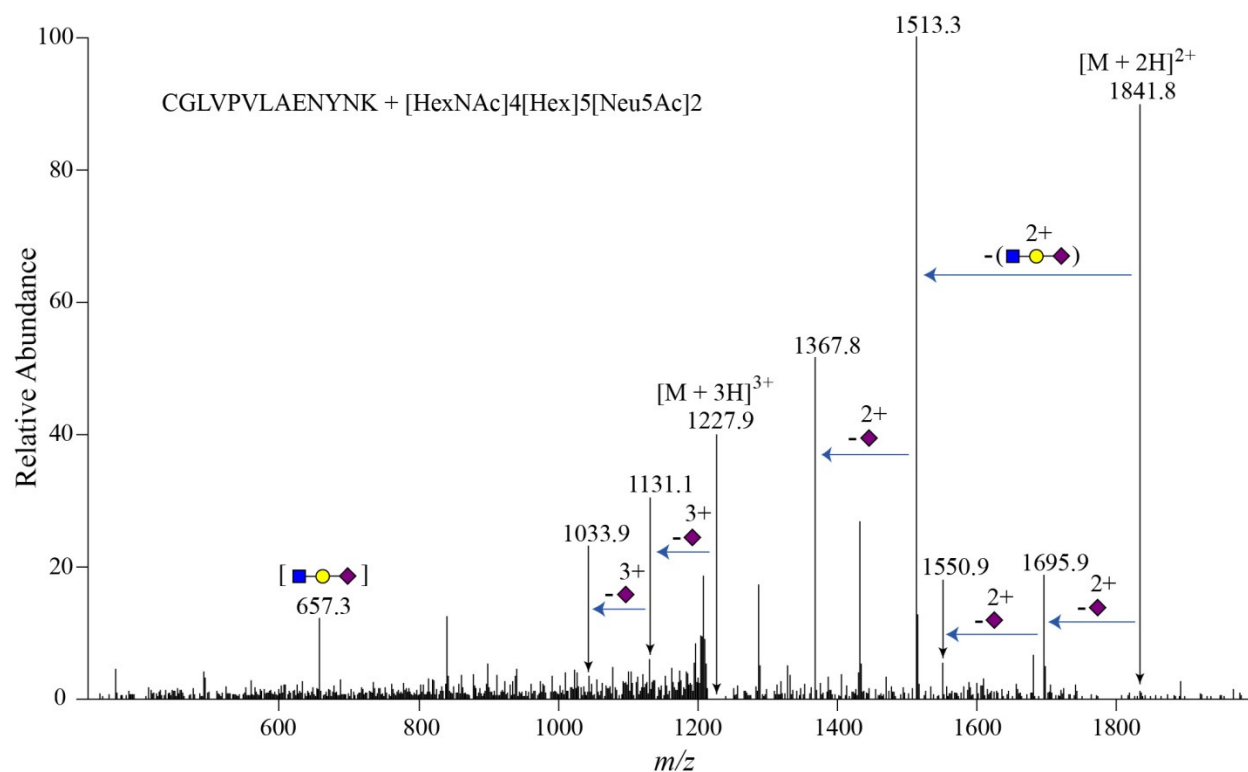


**Figure 2.** Illustration of the two different fragmentation pathways observed in CID spectra of sialylated glycopeptides. In (A) the predominant cleavage observed for glycopeptide precursors in the 1 + and 2 + charge states is shown, whereas the most likely point of cleavage detected for glycopeptides 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.<sup>14, 22, 23</sup> 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.<sup>14</sup> However, they report that sialic acid residues with a

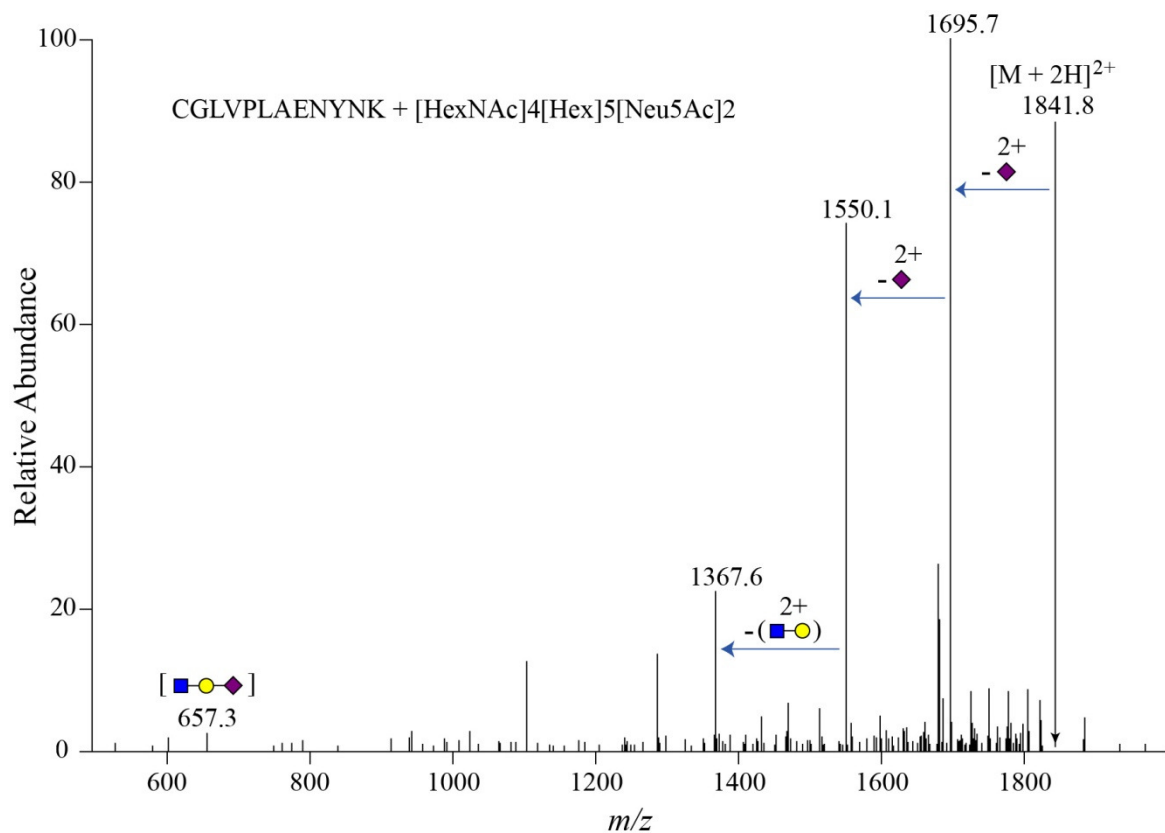
charge state *equal to* 3 + were found to dissociate by the charge-remote cleavage indicative of lower charge state species. In comparison, according to our collection of CID MS/MS data, these 3 + glycopeptide species dissociate by both the charge-remote and charge-directed pathways.

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.<sup>14</sup> 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.<sup>15</sup> 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.

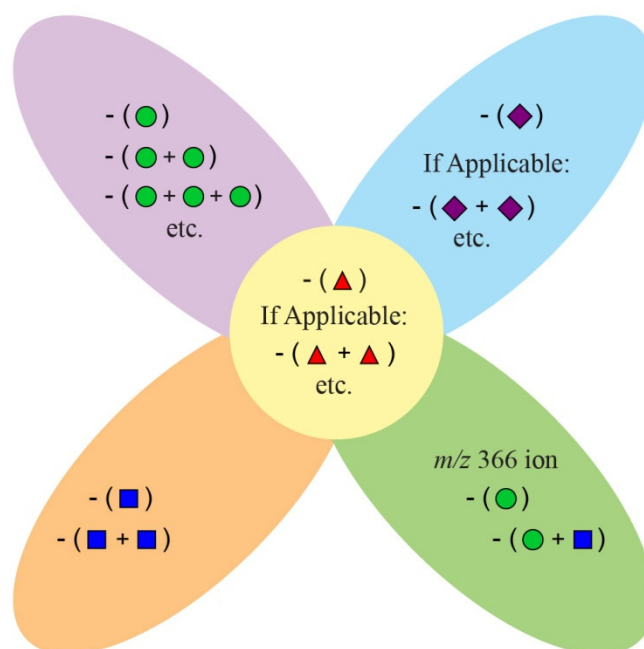
The spectra of the higher charge state species are also shown to contain an intense glycopeptide marker ion at  $m/z$  657, which is also depicted in Figure 3. This marker ion is present in the spectra of most compositions that contain sialic acid, regardless of charge state.<sup>21, 24, 25, 26</sup> 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.<sup>27</sup> 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.<sup>28</sup>

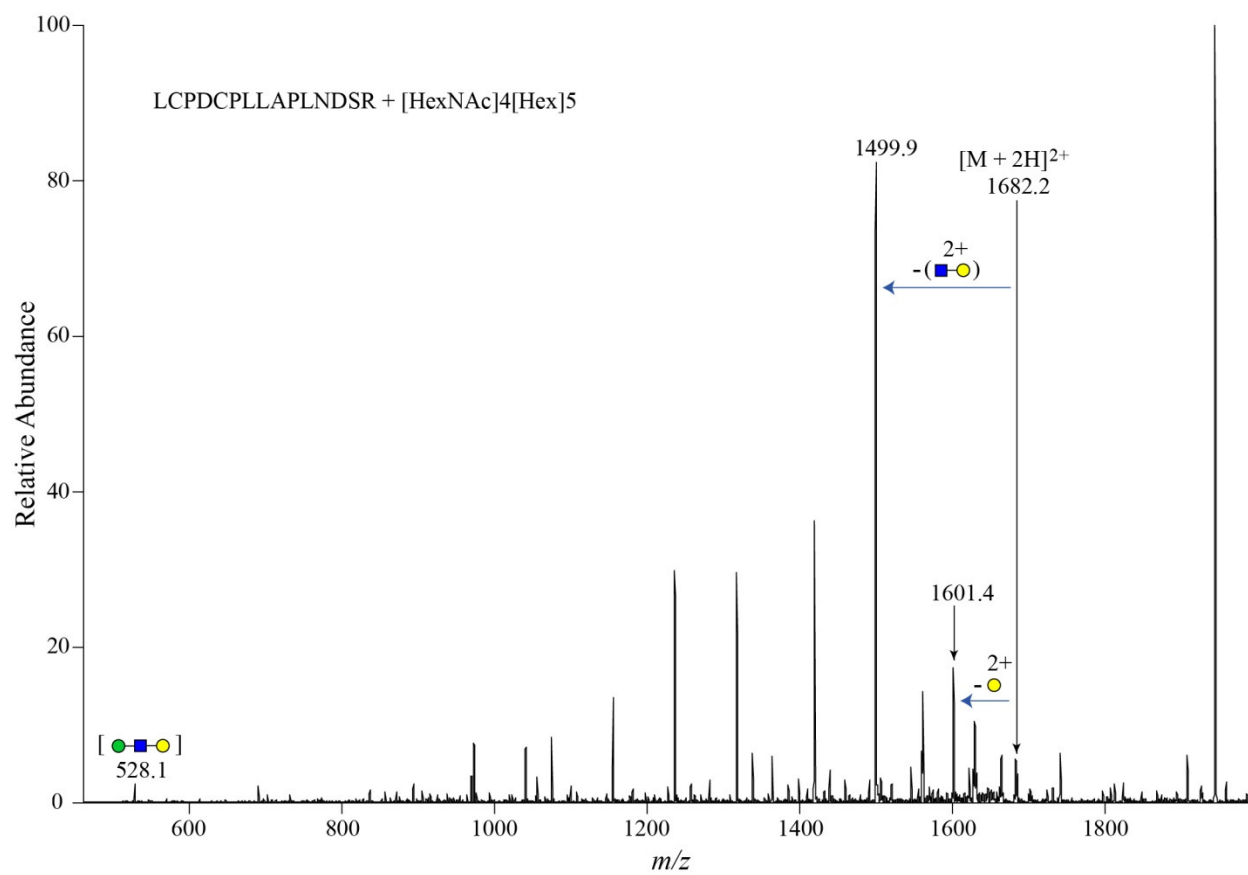
Often, because of the low molecular weight scan cut-off used, the marker ion at  $m/z$  366 is out of scan range when glycopeptide MS/MS data is acquired on tryptic digests of glycoproteins. This occurs because the lowest mass range is limited on an ion trap instrument to approximately 1/3 of the precursor ion's  $m/z$ .<sup>29</sup> 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.

## [Precursor – Monosaccharide] Product Ions Currently Scored by GPG



**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.<sup>15</sup>

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.



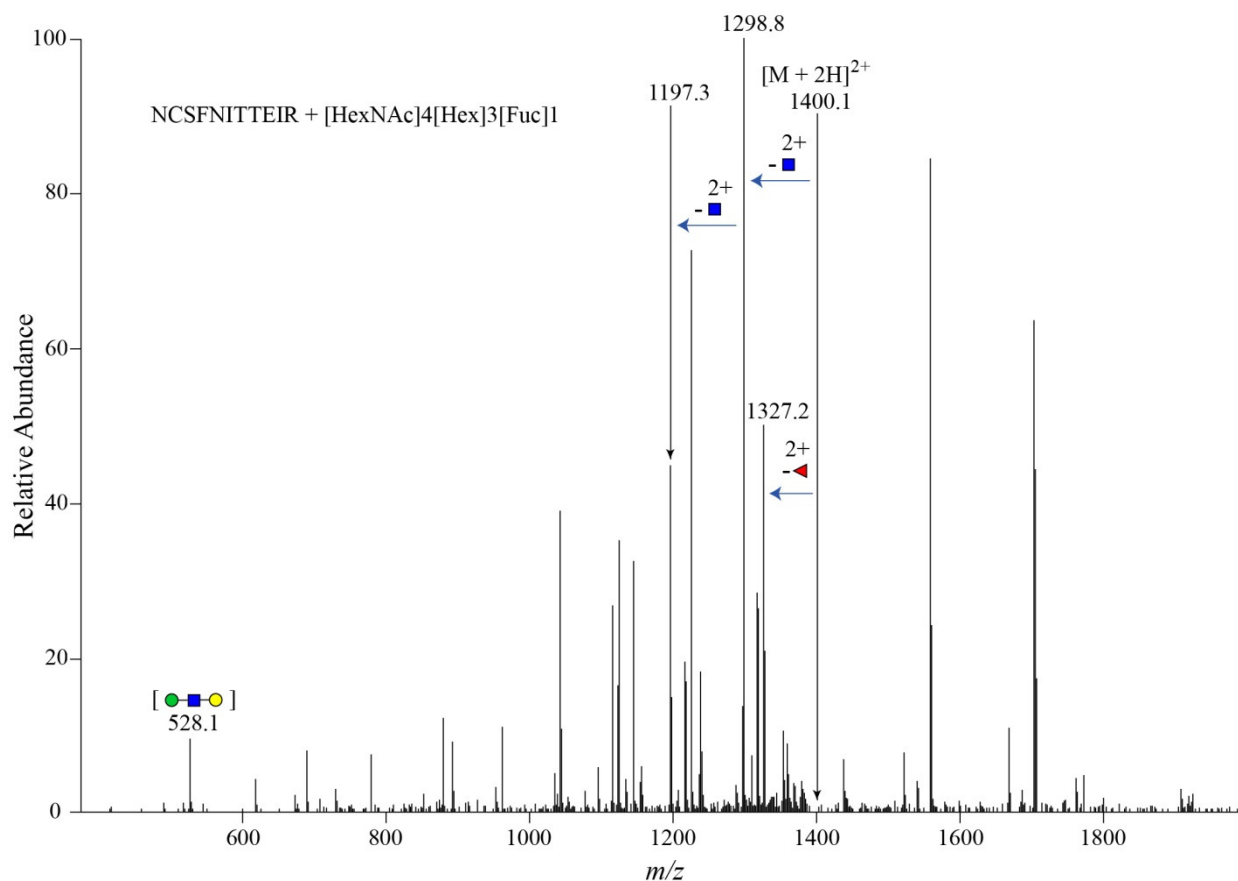
**Figure 6.** CID spectrum from a complex type asialofetuin *N*-linked glycopeptide precursor in the 2+ charge state. An intense marker ion at  $m/z$  528 is shown to be present in the spectrum. The current version of GPG evaluates the presence of a marker ion at  $m/z$  366. This product ion is often 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.

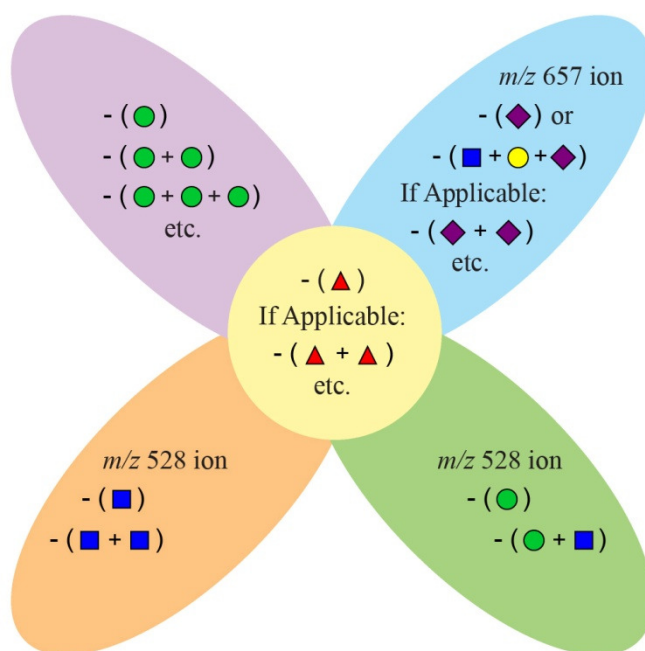
In CID experiments fucose is considered labile, and numerous instances of rearrangement events have also been documented for fucosylated species.<sup>17, 30</sup> For complex/hybrid type glycopeptides appended with fucose, GPG not only evaluates the presence or absence of the product ions formed by the 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.

## [Precursor – Monosaccharide] Product Ions Proposed for Future GPG Scoring



**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.<sup>15</sup>

**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 matrixes.<sup>21, 31, 32</sup> 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

- (1) Mariño, K.; Bones, J.; Kattla, J. J.; Rudd, P. M. A systematic approach to protein glycosylation analysis: A path through the maze. *Nat. Chem. Biol.* **2010**, *6*, 713-723.
- (2) Leymarie, N.; Zaia, J. Effective use of mass spectrometry for glycan and glycopeptide structural analysis. *Anal. Chem.* **2012**, *84*, 3040-3048.
- (3) Nwosu, C. C.; Strum, J. S.; An, H. J.; Lebrilla, C. B. Enhanced detection and identification of glycopeptides in negative ion mode mass spectrometry. *Anal. Chem.* **2010**, *82*, 9654-9662.
- (4) Nie, H.; Li, Y.; Sun, X. L. Recent advances in sialic acid-focused glycomics. *J. Proteomics.* **2012**, *75*, 3098-3112.
- (5) Qiu, R.; Regnier, F. E. Comparative glycoproteomics of *N*-linked complex-type glycoforms containing sialic acid in human serum. *Anal. Chem.* **2005**, *77*, 7225-7231.
- (6) Byrne, B.; Donohoe, G. G.; O'Kennedy, R. Sialic acids: Carbohydrate moieties that influence the biological and physical properties of biopharmaceutical proteins and living cells. *Drug Discov. Today.* **2007**, *12*, 319-326.
- (7) Schauer, R. Sialic acids as regulators of molecular and cellular interactions. *Curr. Opin. Struct. Biol.* **2009**, *19*, 507-514.
- (8) Varki, A. Sialic acids in human health and disease. *Trends Mol. Med.* **2008**, *14*, 351-360.
- (9) Li, H.; d'Anjou, M. Pharmacological significance of glycosylation in therapeutic proteins. *Curr. Opin. Biotechnol.* **2009**, *20*, 678-684.
- (10) Sethuraman, N.; Stadheim, T. A. Challenges in therapeutic glycoprotein production. *Curr. Opin. Biotechnol.* **2006**, *17*, 341-346.
- (11) Bork, K.; Horstkorte, R.; Weidemann, W. Increasing the sialylation of therapeutic glycoproteins: The potential of the sialic acid biosynthetic pathway. *J. Pharm. Sci.* **2009**, *98*, 3499-3508.
- (12) Salinas, P. A.; Miller, M. J. C.; Lin, M. X.; Savickas, P. J.; Thomas, J. J. Mass spectrometry-based characterization of acidic glycans on protein therapeutics. *Int. J. Mass Spectrom.* **2012**, *312*, 122-134.
- (13) Raju, T. S.; Briggs, J. B.; Chamow, S. M.; Winkler, M. E.; Jones, A. J. S. Glycoengineering of therapeutic glycoproteins: In vitro galactosylation and sialylation of glycoproteins with terminal *N*-acetylglucosamine and galactose residues. *Biochemistry.* **2001**, *40*, 8868-8876.
- (14) Zhang, Z.; Shah, B. Prediction of collision-induced dissociation spectra of common *N*-glycopeptides for glycoform identification. *Anal. Chem.* **2010**, *82*, 10194-10202.

- (15) Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. GlycoPep Grader: A web-based utility for assigning the composition of *N*-linked glycopeptides. *Anal. Chem.* **2012**, *84*, 4821-4829.
- (16) Desaire, H.; Hua, D. When can glycopeptides be assigned based solely on high-resolution mass spectrometry data? *Int. J. Mass. Spectrom.* **2009**, *287*, 21-26.
- (17) Tajiri, M.; Kadoya, M.; Wade, Y. Dissociation profile of protonated fucosyl glycopeptides and quantitation of fucosylation levels of glycoproteins by mass spectrometry. *J. Proteome Res.* **2009**, *8*, 688-693.
- (18) Seipert, R. R.; Dodds, E. D.; Clowers, B. H.; Beecroft, S. M.; German, J. B.; Lebrilla, C. B. Factors that influence fragmentation behavior of *N*-linked glycopeptide ions. *Anal. Chem.* **2008**, *80*, 3684-3692.
- (19) Rebecchi, K.R.; Wenke, J. L.; Go, E.P.; Desaire, H. Label-free quantitation: A new glycoproteomics approach. *J. Am. Soc. Mass. Spectrom.* **2009**, *20*, 1048-1059.
- (20) Go, E. P.; Irungu, J.; Zhang, Y.; Dalpathado, D. S.; Liao, H. X.; Sutherland, L. L.; Alam, S. M.; Haynes, B. F.; Desaire, H. Glycosylation site-specific analysis of HIV envelope proteins (JR-FL and CON-S) reveals major differences in glycosylation site occupancy, glycoform profiles, and antigenic epitopes' accessibility. *J. Proteome Res.* **2008**, *7*, 1660-1674.
- (21) Brown, K. J.; Vanderver, A.; Hoffman, E. P.; Schiffman, R.; Hathout, Y. Characterization of transferrin glycopeptide structures in human cerebrospinal fluid. *Int. J. Mass Spectrom.* **2012**, *312*, 97-106.
- (22) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76*, 3908-3922.
- (23) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* **2005**, *77*, 6364-6373.
- (24) Imre, T.; Schlosser, G.; Pocsfalvi, G.; Siciliano, R.; Molnár-Szöllösi, É.; Kremmer, T.; Malorni, A.; Vékey, K. Glycosylation site analysis of human alpha-1-acid glycoprotein (AGP) by capillary liquid chromatography-electrospray mass spectrometry. *J. Mass Spectrom.* **2005**, *40*, 1472-1483.
- (25) Song, E.; Pyreddy, S.; Mechref, Y. Quantification of glycopeptides by multiple reaction monitoring liquid chromatography/tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 1941-1954.
- (26) Chen, G.; Pramanik, B. N. Application of LC/MS to proteomics studies: Current status and future prospects. *Drug Discov. Today.* **2009**, *14*, 465-471.
- (27) Huddleston, M. J.; Bean, M. F.; Carr, S. A. Collisional fragmentation of glycopeptides by

electrospray ionization LC/MS and LC MS/MS: Methods for selective detection of glycopeptides in protein digests. *Anal. Chem.* **1993**, *65*, 877-884.

(28) Conboy, J. J.; Henion, J. D. The determination of glycopeptides by liquid-chromatography mass-spectrometry with collision-induced dissociation. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 804-814.

(29) Lill, J. Proteomic tools for quantitation by mass spectrometry. *Mass Spectrom. Rev.* **2003**, *22*, 182-194.

(30) Froehlich, J. W.; Barboza, M.; Chu, C.; Lerno, L. A.; Clowers, B. H.; Zivkovic, A. M.; German, J. B.; Lebrilla, C. B. Nano-LC-MS/MS of glycopeptides produced by nonspecific proteolysis enables rapid and extensive site-specific glycosylation determination. *Anal. Chem.* **2011**, *83*, 5541-5547.

(31) Zhao, J.; Qiu, W.; Simeone, D. M.; Lubman, D. M. *N*-linked glycosylation profiling of pancreatic cancer serum using capillary liquid phase separation coupled with mass spectrometric analysis. *J. Proteome Res.* **2007**, *6*, 1126-1138.

(32) Jiang, H.; Desaire, H.; Butnev, V. Y.; Bousfield, G. R. Glycoprotein profiling by electrospray mass spectrometry. *J. Am. Soc. Mass. Spectrom.* **2004**, *15*, 750-758.



## 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.<sup>1,2,3</sup> The characterization of these and other PTMs are routinely performed using mass spectrometry (MS) and tandem mass spectrometry (MS/MS) experiments,<sup>4,5</sup> and the availability of reliable automated tools greatly increases the amount of data that can be processed in a given amount of time.<sup>6</sup>

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.<sup>6,7</sup> To overcome this limitation, we developed the GlycoPep Grader (GPG) program to automate the compositional determination of *N*-linked glycopeptides from CID spectra.<sup>7</sup> 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.<sup>7</sup>

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.<sup>8</sup> 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

- (1) Sato, Y.; Inaba, K. Disulfide bond formation network in the three biological kingdoms, bacteria, fungi and mammals. *Febs J.* **2012**, *279*, 2262-2271.
- (2) Jensen, O. N. Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 391-403.
- (3) Walsh, C. T.; Garneau-Tsodikova, S.; Gatto, G. J. Protein posttranslational modifications: The chemistry of proteome diversifications. *Angew. Chem. Int. Edit.* **2005**, *44*, 7342-7372.
- (4) Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21*, 255-261.
- (5) Dalpathado, D. S.; Desaire, H. Glycopeptide analysis by mass spectrometry. *Analyst.* **2008**, *133*, 731-738.
- (6) Woodin, C. L.; Maxon, M.; Desaire, H. Software for automated interpretation of mass spectrometry data from glycans and glycopeptides. *Analyst.* **2013**, *138*, 2793-2803.
- (7) Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. GlycoPep Grader: A web-based utility for assigning the composition of *N*-linked glycopeptides. *Anal. Chem.* **2012**, *84*, 4821-4829.
- (8) Sharma, V.; Eng, J. K.; Feldman, S.; von Haller, P. D.; MacCoss, M. J.; Noble, W. S. Precursor charge state prediction for electron transfer dissociation tandem mass spectra. *J. Proteome Res.* **2010**, *9*, 5438-5444.