

SITE-SPECIFIC ANALYSIS OF GLYCOSYLATED PROTEINS USING MASS SPECTROMETRY

By

Janet W. Irungu

PhD., University of Kansas 2007

Submitted to the Department of Chemistry and the
Faculty of the Graduate School of the University of Kansas
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Heather Desaire (Chairperson)

Mario Rivera

Cindy Berrie

Peter Hierl

Rick Dobrowsky

Date defended

The Dissertation Committee for Janet W. Irungu certifies that this is the approved version of the following dissertation:

**SITE-SPECIFIC ANALYSIS OF GLYCOSYLATED
PROTEINS USING MASS SPECTROMETRY**

Committee:

Heather Desaire (Chairperson)

Mario Rivera

Cindy L. Berrie

Peter Hierl

Rick Dobrowsky

Date Approved _____

Abstract

Site-specific analysis of glycosylated proteins using mass spectrometry

By

Janet W Irungu

Doctor of Philosophy in Chemistry

University of Kansas

Among the numerous post-translational modifications that a protein can undergo, glycosylation is by far the most common, having the most profound influence on the structural and functional properties of the protein. Therefore, profiling glycosylation patterns in glycosylated proteins and defining the structures and locations of these glycans is important in understanding the structure-function relationship of glycans in glycosylated proteins. The work presented herein focuses on applying different mass-spectrometric methods to profile glycosylation patterns in glycoprotein hormones and HIV envelope proteins. To determine the structures and locations of the glycans on these proteins, a glycopeptide-based mass mapping approach was employed.

Glycoprotein hormones mainly contain acidic glycans that are highly sulfated and/or sialylated. These acidic functional groups affect the biological clearance of these proteins. To characterize the glycan structures on

glycoprotein hormones, we used a non-specific enzyme to generate small glycopeptides that are easier to separate and analyze. However, analysis of these glycopeptides can be challenging since it involves simultaneous analysis of two unknowns; the peptide and the glycan portions.

To facilitate identification of the peptide portion, we developed a web-based tool (GlycoPep ID) that utilizes a characteristic product ion observed in (-) MS/MS data of these glycopeptides. To identify the glycan portion, since (-) MS/MS analysis gives very minimal glycan structural information; we developed an ion-pairing approach, which provides a wealth of structural information on the glycan portion of these glycopeptides.

Finally, an HIV envelope protein, CON-S gp140 Δ CFI, a potential vaccine candidate for HIV/AIDS, was characterized. This protein is extensively glycosylated with over 50% of its mass constituting of glycans. Although these glycans play a major role in viral defense mechanism against the host immune system, the structures and locations of these glycans are still not yet known. To develop an efficacious vaccine against this virus, a complete characterization of these glycans is required. A full glycosylation site-specific analysis of glycans in this protein was performed. This information provided biological insights into why CON-S gp140 Δ CFI is a good immunogen, thus a potential candidate for an HIV vaccine.

Acknowledgments

First, I would like to thank my wonderful family for all their love, prayers, and support in every aspect of my life, and though far, you are all so close and dear to my heart. You are the best family anyone could ever wish for and I am so blessed to have each one of you in my life. My heart always overflows with joy whenever I think of each one of you and I could never have made it this far without having you all in my life. I love you all and may God richly bless you in everyway. Mum and Dad, you are a fountain of wisdom, my inspiration, and I have always looked up to you. Thank you for believing in me, for being so encouraging and for the many words of wisdom that you have given me throughout my life. You taught me how to keep focused on the goal and never to allow circumstances around me to distract me from achieving my goals in life. Mum thank you for teaching me how to patiently endure without quitting and Dad I always remember your words, “gutumanaga gugikia” “it’s darkest just before morning”, meaning it’s tougher towards the end. It has been a long night, but I thank God the morning is finally here. It is the faith in God and the values that you both instilled in me that has kept me going.

Secondly, I would like to thank all the people who have supported me in my studies at KU. My advisor, Dr Heather Desaire, I thank you very much for all your advice, help and guidance throughout my doctoral studies. I am

particularly grateful for all the times that you went out of your way to help even when you did not have to and your willingness to help in both research related and unrelated issues. I am also grateful for sending me to conferences through which I got the opportunity to interact and learn from other people in my field. I would also like to thank all the members in Heather's research group, both current and past. I am very grateful to each one of you for all the endless support and help that you gave me throughout the years that we worked together. It was nice working with each one of you and though I cannot mention you all by name I hope you all know that I truly appreciate each one of you. Hui and Eden, thank you for all your help and guidance in my research work. Hui your humble spirit touched my life. Eden you have not only been a mentor in research but spiritually too. I particularly admire your devotion to God and your family. I cannot forget Dr. Todd Williams. Thank you Uncle Todd for all the countless times you helped me with the mass spec instruments. My sincere gratitude for trusting me with your HPLC columns, and even going out of your way to buy for me more columns and the nano-spray tips. You never had to do any of those things but you did them anyway and I appreciate each one of them. I would also like to thank my former supervisors at Midwest Research Institute (MRI), Dr. Joe Algaier and Dr. Pete Schebler. Thank you for all your encouraging words and advice especially the first two years of my graduate studies. Many thanks to Dr. Patricia Chernovitz from Park University for seeing the best in me and

believing in me. I am forever grateful to you for helping me discover my potential in research. I would also like to thank all my committee members for their time, effort and support during the major transitions of my graduate work.

Many thanks to all my wonderful friends, both near and far, for all their moral support, prayers and encouragement. I am so thankful to God for bringing each one of you into my life. I may not be able to thank each one of you in person but I pray that God will reward each one of you a thousand fold for all the many ways you have brought joy in to my life. Rose I could never ask for a better friend. Thank you for being so supportive and for being there for me throughout the many years we have shared together. Everyone should have a friend like you and I am very grateful to God that I have you. I am also very thankful to my sister Ciru. You have not only been a sister to me but my best friend and a source of comfort to me. I thank God every time I think of you and all the things we have shared and been through together.

Finally, I would like to thank my soul mate, my very best friend, Harrison, for always being a blessing to me. You mean the world to me and you are all I could ever ask for or imagine in a friend. Thank you for being so loving, understanding and supportive in all aspects of my life. Thank you for always seeing the best in me and for standing by my side in both good and tough times. Your words “together we can weather any storm in God’s trust” always strengthens my heart. I am looking forward to the rest of our lives

together and I hope to bring as much joy to your life as you have brought in mine. And to the family that we will have together, this work is dedicated to you. Above all, I thank God for everything that he has blessed me with and for bringing me this far. Having you Lord in my life gives me such tremendous joy, such hope and always gives me the strength to carry on. Through it all, I have learned to rejoice in my weaknesses because it is during those times that your power is made perfect.

Table of Contents

Chapter 1: Introduction

1.1 Mass spectrometer.....	1
1.1.1 Ionization source.	2
1.1.2 Electrospray ionization.....	3
1.1.2.1 Mechanism of ESI.....	5
1.1.3 Mass analyzers.....	8
1.2 Fourier transfer mass spectrometry.....	9
1.2.1 Instrumentation.....	10
1.2.2 General principle of ion cyclotron resonance (ICR).....	12
1.2.3 Ion injection into ICR.....	14
1.2.4 Ion trapping in the ICR.....	16
1.2.5 Ions excitation and detection in FTICR.....	17
1.3 MS/MS experiments.....	18
1.3.1 MS ⁿ experiments in ICR CELLS.....	19
1.4 The role of mass spectrometry in glycoprotein analysis.....	20
1.5 Glycosylated proteins.....	22
1.5.1 Biosynthesis of N-glycosylated proteins.....	23
1.5.2 Factors affecting biosynthesis of N-glycosylated proteins..	26
1.5.3 Biological importance of glycosylated proteins.....	27
1.6 An overview and summary of the following chapters.....	28

1.7 References.....	34
---------------------	----

Chapter 2: A Method for characterizing sulfated glycoproteins in a glycosylation site-specific fashion, using ion pairing and tandem mass spectrometry

2.1 Introduction.....	41
2.2 Experimental.....	45
2.2.1 Digestion of eTSH with Proteinase K.....	45
2.2.2 Peptide Sequencing (Edman).....	45
2.2.3 eTSH glycopeptide preparation for MS analysis.....	45
2.2.4 Mass Spectrometry	46
2.2.5 MS/MS analysis	47
2.3 Results and Discussion.....	47
2.3.1 Compositional analysis of eTSH glycopeptides	47
2.3.2 Comparison of (-)ESI-MS/MS data	50
2.3.2.1 Number of SO ₃ groups present	50
2.3.2.2 Different glycosylation sites.....	52
2.3.2.3 Different charge states	52
2.3.3 Useful structural information based on fragmentation characteristics observed in (-)ESI-MS/MS.....	57
2.3.4 New Approach: The use of ion-pairing with MS/MS.....	58
2.3.5 Comparison of MS/MS data of ion-pair complexes	62

2.3.5.1 Number of SO ₃ groups present.....	62
2.3.5.2 Different glycosylation sites.....	65
2.3.6 Application to unknown glycopeptides	68
2.3.6.1 Example 1: Use of MS/MS to differentiate two isobaric structures.....	68
2.3.6.2 Example 2: Use of MS/MS analysis for structural information.....	74
2.4 Conclusion.....	76
2.5 References.....	77

Chapter 3: Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID

3.1 Introduction.....	80
3.2 Experimental.....	86
3.2.1 Enzymatic digestion of eTSH and eFSH	86
3.2.2 Glycopeptides preparation for mass spectrometry analysis.....	86
3.2.3 CID experiments.....	86
3.2.4 Data analysis	87
3.3 Results and Discussion.....	88
3.3.1 GlycoPep ID Overview.....	88

3.3.2 Method validation.....	92
3.3.3 Application to complex negatively charged glycopeptides from eFSH.....	101
3.3.4 CID experiment data for glycopeptides containing sialic acid only.....	101
3.3.5 CID experiments for glycopeptides containing both sulfate and sialic acid.....	103
3.4 Conclusion.....	105
3.5 References.....	108

**Chapter 4: Comparison of LC/ESI-FTICR MS vs MALDI-TOF/TOF for
glycopeptide analysis of highly glycosylated protein: HIV Envelope
glycoprotein**

4.1 Introduction.....	111
4.2 Experimental section.....	115
4.2.1 Materials and Reagents.....	115
4.2.2 Trypsin digestion of CON-S gp140ΔCFI protein.....	116
4.2.3 Reverse phase HPLC fractionation.....	116
4.2.4 Deglycosylation.....	117
4.2.5 MALDI-TOF/TOF MS analysis.....	117
4.2.6 Capillary LC/ESI-FTICR MS analysis.....	118

4.2.7 CID Experiments in LC/ESI-FTICR MS.....	119
4.2.8 Data analysis.....	120
4.3 Results and Discussion.....	121
4.3.1 Assigning glycopeptide compositions	122
4.3.1.1 MS/MS data from LC/ESI-FTICR MS.....	126
4.3.1.2 MS/MS data from MALDI-TOF/TOF MS	127
4.3.2 Number of glycoforms identified	129
4.3.3 Identification of the most abundant type of N-linked glycan present.....	131
4.3.4 Glycosylation sites coverage.....	138
4.4 Conclusion.....	142
4.5 Significance of analyzing CON-S gp140 Δ CFI.....	154
4.5.1 CON-S consensus gene design.....	155
4.5.2 Expression of recombinant HIV envelopes.....	157
4.6 References.....	157
 Chapter 5: Conclusion and Future directions.....	 161

Page left blank intentionally

CHAPTER 1

Introduction

1.1 Mass spectrometry

Mass spectrometry (MS) is an analytical technique that is used to determine the molecular weight of a molecule. Specifically, MS is used to measure the mass-to-charge ratio of a molecule by analyzing its gas phase ions. This is typically done by making ions from the sample molecules and measuring the mass-to-charge ratio (m/z) of each component present in the sample, generating a spectrum that shows the relative abundance of each component according to its m/z ratio.^{1, 2} MS can be used in both qualitative and quantitative studies.

A typical mass spectrometer consists of three basic components; An ionization source, a mass analyzer and a detector (See Figure1-1 below). The analyte is usually introduced into the ionization source through an inlet device, and once in the ion source, ions are generated by inducing loss or gain of a charge from the sample molecule. Ions are then transferred into the mass analyzer where they are separated according to their m/z ratio and counted by the detector.¹

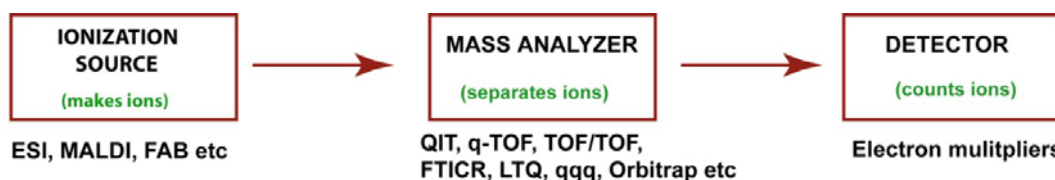


Figure 1-1: A block diagram showing the three basic components of a mass spectrometer: an ionization source, a mass analyzer, and a detector.

1.1.1 Ionization sources

As the name suggests, an ionization source is used to ionize the sample molecule thus generating gas phase ions. This is the most important part of a mass spectrometer.³ Since the invention of mass spectrometry, several ionization sources have been introduced. The most common ones include electron ionization (EI), chemical ionization (CI), fast atom/ion bombardment (FAB), matrix assisted laser desorption ionization (MALDI) and electrospray ionization (ESI). These ionization sources function by ionizing the neutral molecule through electron ejection, electron capture, protonation, deprotonation, adduct formation, cationization, or through the transfer of a charged species from condensed phase to gas-phase.^{1, 2}

The mode of ionization selected mainly depends on the physicochemical properties of the analyte. For instance, if the analyte is volatile and thermally stable, ionization techniques like EI and CI would be the most appropriate, since they are very energetic and are more suitable for gas-phase ionization. On the contrary, if the sample is non-volatile and

thermally labile, softer ionization techniques that are capable of generating ions directly from condensed phase to gas-phase ions like MALDI or ESI are more appropriate. MALDI and ESI generate gas-phase ions directly from solid and liquid phase, respectively.^{1, 2, 4} A detailed explanation of how ESI works ensues as an example of how these ionization methods work. A description of how each of the other ionization methods work can be found in references (1,2,4).

1.1.2 Electrospray Ionization (ESI)

The first ESI ionization source was invented in 1989 by John Fenn, when he demonstrated the formation of multiply charged ions from large-molecules, like proteins, which enabled analysis of these molecules with mass spectrometers with limited mass range.⁵ The advent of ESI made tremendous contribution in structural analysis of important biomolecules, since its evaporation ionization process minimizes dissociation of molecular ions during MS experiments.⁶⁻⁸ ESI ionization technique is the softest ionization technique known thus far, and its ability to generate gas phase ions directly from liquid phase has also broadened its applicability, since it can easily be coupled with separation techniques.⁹ As a result, ESI is a choice method for analysis of wide range of compounds (both small and large) and is especially useful during analysis of large nonvolatile biomolecules like proteins, oligosaccharides, and glycoproteins.

ESI ionization process takes place at atmospheric pressure. Figure 1-2 shows how this process occurs. As illustrated in this figure, the sample solution in a suitable solvent mixture is introduced continuously through a capillary tube that is held at a higher potential than the instrument orifice, producing a fine spray of highly charged droplets.¹⁰ The solvent is then evaporated from the charged droplets converting them into gas-phase ions. Once formed, these ions are then driven electro-statically towards the instrument orifice. The ions are then transported to the high-vacuum mass analyzer through a series of pressure-reduction stages. For optimum operation, a normal ESI source typically requires flow rates of 2 –10 $\mu\text{L}/\text{min}$ and can be operated in the negative or positive ion mode by varying the polarity of the voltage applied to the capillary tube.^{1, 4, 10, 11}

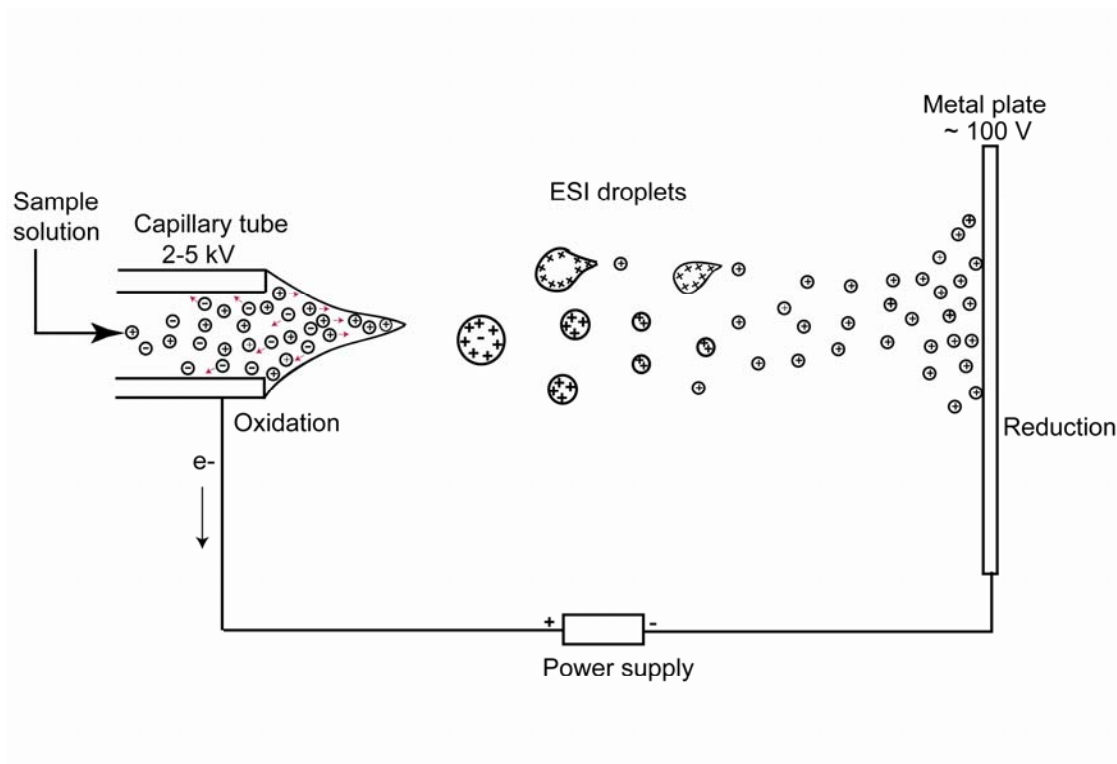


Figure 1-2: A schematic representation of ESI process. (Adapted from Ikonou, Blades and Kebarle.)¹²

1.1.2.1 Mechanism of ESI

Three different processes are involved in ESI ionization; droplet formation, droplet shrinkage, and desorption of gaseous ions.¹⁰ When the sample liquid flows through the capillary tube that is held at a high voltage (2-5 kV), it experiences a strong electric field that causes generation of charges through a redox reaction. For example, if the voltage applied to the capillary is positive, an oxidation reaction occurs in solution at the metal contact of the sample solution whereas a reduction reaction occurs at the

counter electrode. In this case, the continuous removal of negative ions from the metal capillary leads to creation of positive ions. This electrochemical redox reaction is responsible for facilitating the continuous production of charged ions.^{4, 10, 12-14} As a result, the positively charged ions concentrate at the tip of the capillary and are drawn towards the instrument orifice (counter electrode), whereas the anions migrate towards the capillary walls away from the tip, as shown in Figure 1-2. Eventually, the liquid droplet, populated mainly by positive ions, protrudes from the capillary tip in what is known as a “Taylor cone”. When the Coulombic repulsion forces at the surface of the liquid droplet exceeds the Rayleigh limit, a point at which the Coulombic repulsion forces equals the surface tension of the liquid, the liquid droplet explodes into smaller droplets containing an excess of positive charges as shown in Figure 1-2.¹⁵

Once the charged droplet is formed, evaporation of the solvent is attained through application of heated nitrogen causing the droplet to shrink in size. From here, a cascade of ruptures ensues. The charge density on the droplet surface increases as it reduces in size (shrinks) and once again reaching the Rayleigh limit causing fission of the droplets into smaller highly charged droplets. As the solvent evaporation continues, this process occurs repeatedly producing smaller and smaller droplets. There are two mechanisms that have been proposed to explain how the charged droplets produce gas-phase ions, the charge-residue model (CRM) and ion-

desorption model (IDM). CRM proposes that a series of several solvent evaporation and droplet fission occurs repeatedly until a very small droplet containing only one solute molecule is formed.⁴ See Figure 1-3. Eventually all the solvent is evaporated resulting in a single molecule that retains the charge of the droplet, as shown in Figure 1-3(a) This mechanism is attributed to ionization of hydrophilic species. The other mechanism, IDM, proposes that as desolvation continues, the electric field on the surface of the droplet becomes large enough such that direct emission of single ions from the surface occurs. (See Figure 1-3(b)) This mechanism is believed to occur in hydrophobic molecules such as peptides and fatty acids.¹⁰

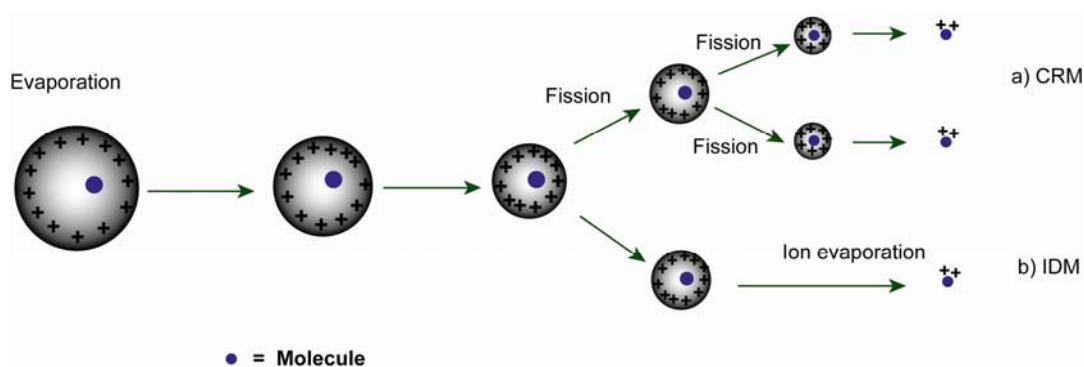


Figure 1-3: A schematic diagram demonstrating the two mechanisms proposed for ESI. The solvent surrounding the molecule is evaporated leading to the formation of a charged molecule. (a) Shows the CRM mechanism; (b) Represents IDM mechanism. (Adapted from Chhabil Dass.⁴)

1.1.3 Mass analyzers

Mass analyzers separate ions according to their mass-to-charge ratio. Different mass analyzers perform this function by separating ions either in time or in space.^{1,2} For example, time of flight (TOF) and magnetic sectors are spatial mass analyzers, whereas trapping mass analyzers like quadrupole ion trap (QIT) and Fourier transform ion cyclotron resonance (FTICR) separate ions in time. There are several characteristics of mass analyzers that are used to determine their performance. These include mass range, accuracy, resolution, MS/MS capabilities, and scan speed.¹ The mass range determines the lowest and the upper m/z that can be measured by the mass analyzer. For example, the mass range for TOF is theoretically unlimited whereas for a QIT, the mass range is up to m/z 3000.¹ The ability for a mass analyzer to separate different m/z accurately and be able to explicitly discriminate them (resolution) also largely determines its performance. Among all the mass analyzers currently available, FTICR has the highest mass accuracy and resolution capabilities.

In addition to mass accuracy and resolution, another important feature of mass analyzers is their ability to perform tandem mass spectrometry experiments. Tandem mass spectrometry, or MS/MS, refers to the ability of the analyzer to isolate a primary ion (precursor ion) in the first MS, activate it, fragment it, and analyze the resulting product ions. Mass analyzers that separate ions in space are mainly capable of performing one step of these

experiments since they are limited to the number of mass analyzers that can be combined in series, whereas those analyzers that separate ions in time can perform multiple stages of the MS/MS experiments (MS^n).² The ability to perform MS^n experiments increases the amount of structural information that can be obtained thus increasing the applicability of time-based mass analyzers like FTICR in structural analysis studies of many biomolecules.

Due to the many benefits derived from FTICR, instruments employing this mass analyzer have gained wide applicability, not just as a typical device that can separate masses based on m/z , but also as a mass analyzer that can perform a variety of other unique functions as described below.

1.2 Fourier Transform Mass Spectrometry (FTMS)

FTMS, also known as Fourier transform ion cyclotron resonance (FTICR) mass spectrometry, is widely recognized as a mass spectrometer with the highest resolution and mass accuracy. The ability to provide exact mass measurements and elemental composition assignments is highly desirable not only for small molecules, but also large biomolecules. Virtually any ionization technique can be coupled to FTICR, which increases its applicability. For example, the simultaneous implementation of both ESI and MALDI, the two most useful ionization methods for oligosaccharide

analysis was first performed by FTICR.¹⁶ The ability to couple FTICR with ESI also allows the analysis of very large molecules through the formation of multiply charged ions with high isotopic resolution, thus increasing the upper mass limit for FTICR MS.^{6, 17} This instrument also provides additional benefits like simultaneous detection of all ions in non-destructive manner making it ideal for analysis of minute quantities of samples.¹⁸ The timescale (milliseconds to hours) for FTICR also makes it a versatile instrument that can be used to perform various experiments like slow (and fast) ion-molecule reaction experiments, collision induced dissociation (CID) experiments, photo-dissociation etc.¹⁹ All these benefits collectively make FTICR a very powerful instrument for providing both molecular weight measurements and for structural elucidation studies.

1.2.1 Instrumentation

Since its innovation in 1974 by Comisarow and Marshall, FTICR has undergone a tremendous transformation.²⁰ Currently, there are various designs of FTICR that are commercially available, but regardless of the design, all FTICR share several common features which include a magnet, an analyzer cell, an ultrahigh vacuum system, and a data system.^{18, 21} The best performance for FTICR is achieved with high magnetic field strength and very low pressure.^{22, 23} However, the heart of this instrument is the analyzer cell, which is an ion trap that can act as an ion source, mass

analyzer, and detector, where ions are always separated in time rather than in space.^{18, 24} Because of the pressure differences between an ionization source and a mass analyzer, most FTICR instruments are designed with an external ion source separated spatially from the high vacuum of the mass analyzer cell.^{17, 22} The mass analyzer cell is housed inside a large magnet and can either be cubic or cylindrical in shape. The most common geometry is a cubic cell containing three pairs of electrodes classified as follows; two trapping plates (front and back electrodes), two excitation plates (the side electrodes), and the two detector plates (the top and bottom electrodes).^{18, 21, 24-26} See Figure 1-4. Ions are injected into the analyzer cell along the same direction as the magnetic field lines and they are confined to the center of the cell by application of a small voltage on the trapping plates. The trapped ions can then be manipulated and ultimately detected based on their interactions with the magnetic and electric fields present in the mass analyzer cell.

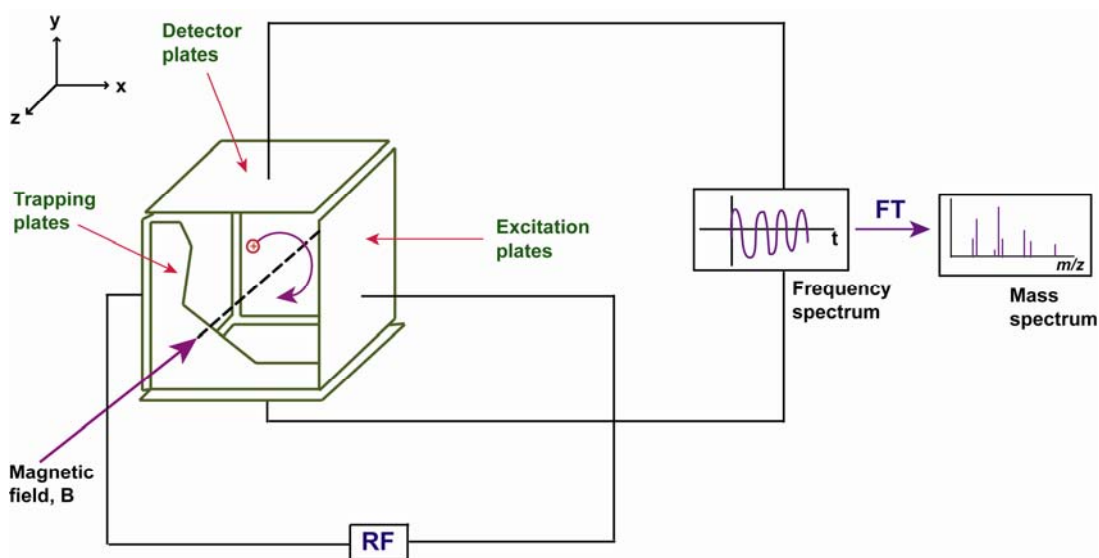


Figure 1-4: Diagram representing FTICR instrument. The mass analyzer comprises of three types of plates, labeled as trapping, excitation and detector plates. Ions enter the ICR cell along the z-direction and rotate along the x-y plane as shown above, inducing an image current that is detected and fourier transformed to give a mass spectrum. (Adapted from <http://www.chm.bris.ac.uk/ms/theory/fticr-massspec.html>, 11/10/07)

1.2.2 General principle of Ion cyclotron resonance (ICR)

An ion of charge (q) and mass (m) travelling at velocity (v) through a uniform static magnetic field (B) experiences a force (F) that is perpendicular to its velocity, causing its motion to curve or rotate perpendicular to the magnetic field direction.^{25, 27-29} See Figure 1-5 below.

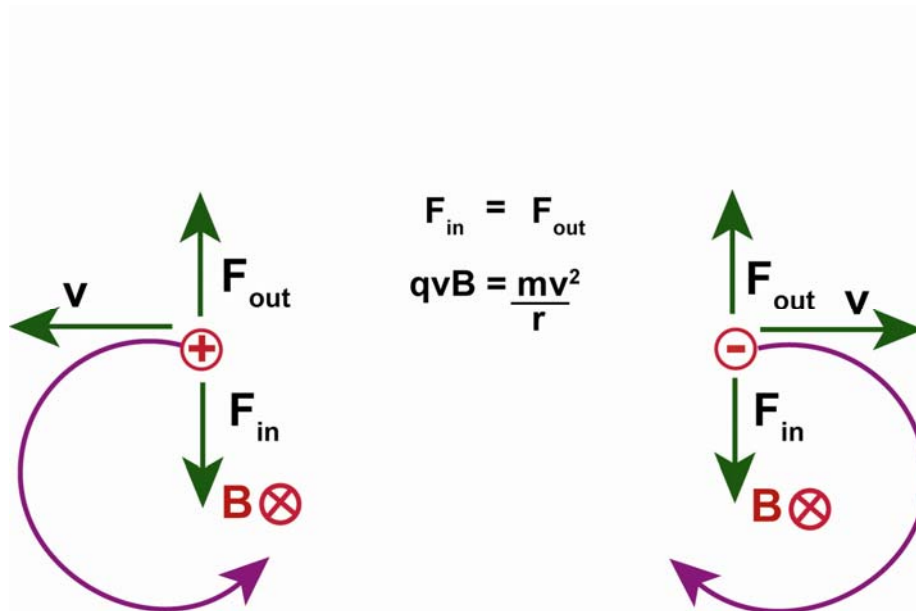


Figure 1-5: Shows the motion of the ion in a magnetic field. The polarity of the ions determines the direction of motion. Figure on the left shows the motion of a positively charged ion and the one on the right indicate the direction of a negatively charged ion in a static magnetic field. (Adapted from Marshall et al.²⁹)

This force (F_{in}), also called Lorentz force is balanced by an outward centrifugal force (F_{out}) stabilizing the ion along its cyclotron path.^{2, 30} See Figure 1-5 above. This can be described by the following basic equations.

$$qv * B = \frac{mv^2}{r}$$

Inward Lorentz force = Outward force

(Equation 1)

Where r is the orbital radius of the ion's motion in the magnetic field. The cyclotron or rotational frequency of the ion can be obtained by rearranging the above equations as follows.

$$\frac{v}{r} = \frac{qB}{m} \quad (\text{Equation 2})$$

Where v/r represents the cyclotron frequency, ω_c , in $2\pi x$ (cycles per second), B in Telsa, m in kilograms, r in meters, q in Coulombs, and v in meters per second.^{25, 28, 30} As shown from equation 2, the cyclotron frequency is inversely proportional to the mass-to-charge ratio (m/q or m/z). All ions of the same m/q rotate at the same cyclotron frequency, independent of their velocity.³⁰ Therefore, the mass-to-charge ratio of an ion is determined by measuring its cyclotron frequency. This is the most notable and unique advantage of FTICR, since frequency can be measured with high precision.^{22, 27} As a result, this technique offers the highest mass resolution and mass accuracy compared to other types of mass measurements.^{30, 31}

1.2.3 Ion injection into ICR

Although there are some ionization techniques like EI that can easily be implemented inside the ICR analyzer cell, it is more beneficial to have the ionization source outside the large magnetic field.^{18, 28, 32} Furthermore, generating ions from nonvolatile molecules can be a daunting task when

performed inside the magnet.^{18, 24} As a result, ions are typically generated by an ionization source located outside the analyzer cell. Several differential pumping stages are placed between the ionization source and the analyzer cell to decouple high pressures of the ion source from the ultra low pressures of the ICR mass analyzer (10^{-9} Torr or less).^{18, 24, 30, 33, 34} However, injecting ions into the analyzer cell is a major hurdle since the ions not only have to penetrate through the large magnetic field, but they also have to slow down in order to be trapped inside the analyzer cell.^{30, 34}

Typically, ions enter the analyzer cell through a small opening on the front trapping plate along the z direction or the magnetic field direction (See Figure 1-4) To allow ions to enter the analyzer cell, the voltage of the front trapping plate is lowered while the one on the back trapping plate is increased to prevent the ions from exiting the analyzer cell.³⁴ The potential of the front trapping plate can then be raised to prohibit ions from returning to the front trapping plate and this also prevents more ions from entering the ion trap (analyzer cell).^{30, 34} To trap ions of a particular m/q (q is the same as z in the term m/z), the potential applied to the trapping plates is usually varied to optimize the trapping of those ions.³⁵ Alternatively, once the ions enter the cell, their kinetic energy can be lowered through collisional dampening with a background gas thus slowing them down and permitting them to be trapped. This allows ions to be accumulated for longer periods.³⁶

1.2.4 Ion trapping in the ICR

Once inside the cell, ions are constrained in the analyzer cell or ion trap by a combination of two forces, electrical and magnetic forces. The main constraining force comes from the magnetic field which “traps” the ions by directing their path in a circular motion away from the analyzer cell walls. This traps the ions in two-dimensions, the x-y plane which is perpendicular to the magnetic field as shown in Figure 1-4. The magnetic field is parallel to the z-axis of the analyzer cell (trap). This field is fixed such that each m/q has a unique cyclotron frequency as described by equation 2. However, this cyclotron equation does not account for the force that constrains the ions in the axial z-direction.^{24, 25, 30, 34} To prevent the ions from escaping along the z-direction, thus trapping them in the third dimension, an electric field is employed. This is done by applying a small voltage (about 1V), of the same polarity as the ions of interest to the trapping plates (both front and back).^{18, 21} By combining the contributions of the magnetic and electric fields, ions orbit with a cyclotron frequency, ω_c , that varies with both magnetic (B) and electric (E_0) field strengths as defined by the following equation.

$$\omega_c = \frac{qB + (q^2B^2 - 4mqE_0)^{1/2}}{2m} \quad \text{(Equation 3)}$$

In the absence of electric field (E_0), equation 3 is equal to equation 2. For further explanation on the basis for this equation, see references (25, 26, 31, and 38)

Other forces that contribute to the outward electric fields include space charge effects. These forces are due to Coulombic repulsion forces between the ions in the trap. The magnitude of these forces increases as the number of ions accumulated in the trap increases. However, the effect of these forces is not as much as the one resulting from the trapping potential.²⁴

1.2.5 Ions excitation and detection in FTICR

Before excitation, the trapped ions move in small cyclotron orbits along or near the z-axis at different cyclotron frequencies depending on their m/z ratio. For instance, ions of the same m/z rotate at the same frequency but occupy different positions about the circular orbit therefore out of phase.^{24, 29} To excite these ions and ultimately detect them, the remaining four plates parallel to the magnetic field are utilized. The excitation process involves the application of a range of radio frequencies (RF) that are in resonance with the ions' cyclotron frequencies to the two excitation plates (see Figure 1-4). Consequently, the ions absorb energy, causing their cyclotron radius to enlarge and become coherent (in-phase); thus, ions of the same m/z follow a coherent orbital path, with the same

radius. If the applied RF voltage is sufficiently high, ions will end up hitting the plates, and they are lost. This method can be used to eject unwanted ions, leaving the ions of interest for further experiments. For detection, the RF is usually turned off before ions can strike the cell plates.^{18, 24}

To detect the ions, an image current that is induced by the ions as they pass by the two detector plates (top and bottom electrodes) is measured. The induced current oscillates at the same frequency as the coherently moving ions, which in turn depends on their m/z values. This current is then detected in an external circuit between the two detection plates. This current is referred to as the image current and is detected as a function of time. Since the ions are not destroyed, they can be re-measured, thus increasing the sensitivity and resolution of this instrument. The current is then amplified, and Fourier transformed from the time-domain to the frequency domain, generating a mass spectrum as shown in Figure 1-4.

1.3 MS/MS experiments

MS/MS experiments, also known as tandem MS, involves separating or isolating the ion of interest (precursor ion) in the first MS, which is then followed by activation and fragmentation of that ion to produce its product ions. Fragmentation of the precursor ion is usually achieved by colliding the ion it with inert gas molecules. The fragmentation/product ions are then

collectively used to provide structural information of the precursor ion. As previously mentioned (see section 1.1.3), the ability to perform MS/MS experiments depends on the instrument (mass analyzer) employed. FTICR is one of the few instruments capable of performing MSⁿ experiments, which are extremely useful for structural elucidation of complex molecules like glycoproteins and oligosaccharides. In addition, FTICR gives the best isotopic resolution for multiply charged ions making it possible to interpret mass spectral data for large molecules ionized by ESI.^{16, 37}

1.3.1 MSⁿ experiments in ICR cells

Once ions are trapped inside an FTICR mass analyzer cell, a precursor ion can be selected for further experiments. The excitation process described previously can be used to eject all unwanted ions from the analyzer cell by causing them to collide with the cell plates. After the ion of interest, or precursor ion, is isolated, various methods can be utilized to activate the precursor ion to promote collision induced dissociate (CID). Sustained off-resonance irradiation (SORI) is the most commonly used method for activation. In this case the precursor ion is excited at a frequency that is slightly off-resonance causing the ion to be in and out of phase. As a result, the ion is excited and de-excited alternately. A pulsed collision gas is then introduced into the cell when the ions are off-resonance excitation.^{18, 21} The low energy collisions are sustained and eventually the

precursor ion gains enough energy to cause it to dissociate. This method is more efficient and selective than other modes of excitation, including on resonance excitation and stored waveform inverse Fourier transform (SWIFT) excitation. A description of the fundamentals of the other excitation methods is provided by Marshall and co-workers.^{21, 27, 38}

Other types of MS/MS experiments that can be performed in FTICR include photo-dissociation of large molecules. These include multi-photon infrared photo-dissociation (IRMPD) where infrared (IR) laser photons are employed to slowly heat the molecules thus producing low-energy fragments similar to CID.³⁹ IRMPD differs from CID in that unlike CID, it does not require the use of gas pulses hence the dissociated ions can be detected at high resolution as they form.⁴⁰ Alternatively, blackbody infrared radiative dissociation (BIRD) can be employed whereby a thermally heated ICR cell is utilized.⁴¹ Another MS/MS technique that can be performed in FTICR is electron-capture dissociation (ECD), which involves exposing multiply charged cations trapped in ICR cells to low-energy electrons.^{17, 42}

1.4 The role of mass spectrometry in glycosylated proteins analysis

Mass spectrometry (MS) is one of the most versatile analytical techniques that has gained widespread use in glycosylated protein analysis, due to its high selectivity and sensitivity and the ability to analyze complex mixtures rapidly. The advent of soft ionization techniques such as fast atom

bombardment (FAB), electrospray ionization (ESI)⁵ and matrix assisted laser desorption ionization (MALDI)⁴³ revolutionized many research fields by providing new insights into the structural details on many levels for various important classes of biomolecules. One of those fields, which has benefited tremendously from MS is the glycoproteomics field. New inventions in MS continue to make enormous contribution in this field.

Glycoprotein analysis by MS is typically achieved by two main approaches labeled as (1) and (2) on the schematic diagram below (Figure 1-6). In approach (1), the glycoprotein is proteolytically digested resulting to a mixture of peptides and glycopeptides, while approach (2) involves cleaving off the glycans from the protein backbone either enzymatically or chemically producing a mixture of peptides and glycans. In both approaches, a separation step is required to remove the peptides prior to MS analysis. Approach (1) is more advantageous than approach (2) since it does not require extra sample manipulation like derivatization,⁴⁴ and allows site-specific glycosylation profiling.⁴⁵

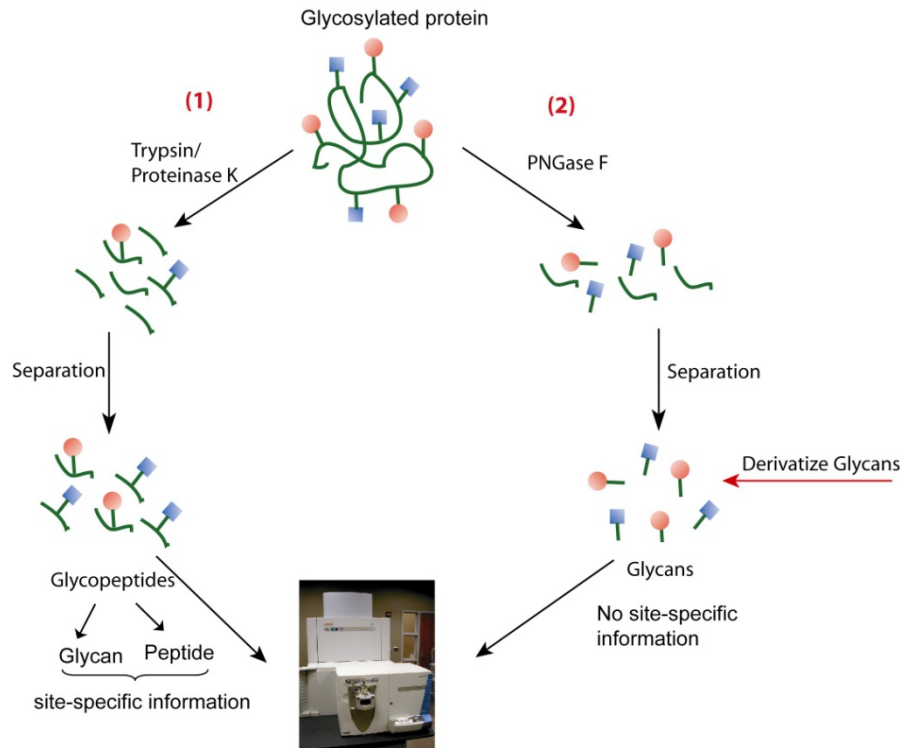


Figure 1-6: A schematic diagram illustrating the two main approaches used in glycosylated proteins analysis by MS. Approach (1) uses either trypsin (specific) or proteinase K (non-specific) enzyme to produce a mixture of peptides and glycopeptides. Approach (2) employs PNGase F (an enzyme that cleaves off N-linked glycans) or chemicals like hydrazine to cleave off the glycans from the protein resulting in a mixture of glycans and peptides.

1.5 Glycosylated proteins

Glycosylated proteins, or glycoproteins, are proteins to which carbohydrates moieties are covalently attached through glycosidic bonds. This process is known as glycosylation and is the most ubiquitous modification that occurs on cellular and secreted proteins.⁴⁶⁻⁴⁸ Although very common, glycosylation is a highly specific process and mainly occurs on three amino acids residues; asparagines (Asn), serine (Ser), and

threonine(Thr). There are only two ways glycans can be attached to these amino acid residues; through the amide nitrogen of Asn forming the N-glycosidic bond; or through the hydroxyl oxygen of Ser or Thr forming the O-glycosidic bond, hence the N- and O-glycosylation types, respectively. N-glycosylation is the most common type of glycosylation in eukaryotic cells and some prokaryotic cells, and is therefore the focus of the studies presented herein.^{49, 50}

1.5.1 Biosynthesis of N-glycosylated proteins

The process by which proteins are N-glycosylated in mammalian systems is highly complex and follows a series of pathways. This process occurs both co- and post-translationally in proteins bearing an asparagine in a consensus sequence of tripeptide Asn-Xaa-Ser/Thr, where X can be any amino acid except for proline, thus each of these tripeptide sequences constitute a potential glycosylation site.⁵⁰ N-glycosylation starts when a preformed precursor oligosaccharide ($\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$) attached on a carrier lipid is transferred to a nascent protein chain (containing the Asn-Xaa-Ser/Thr sequon) in the endoplasmic reticulum (ER) as shown in Figure 1-7. As the nascent protein chain continues to grow, several subsequent reactions occur that are catalyzed by enzymes, (“modifying enzymes 1” shown in Figure 1-7) specifically, glucosidases and mannosidases present in the ER, leading to the precursor oligosaccharide

being trimmed down to a high-mannose N-linked glycan. Further trimming and processing of the precursor oligosaccharide may take place as the protein migrates to the Golgi apparatus (GA) where it encounters various processing enzymes (galactosyl, sialyl, fucosyl transferases, GlcNAc and GalNAc transferases – “modifying enzymes 2” shown in Figure 1-7) to produce hybrid and complex type of N-linked glycans.⁴⁹ This results in N-linked glycans that are highly heterogeneous, particularly in the terminal residues. Thus, structural diversification of N-linked glycans occurs in the GA where the N-linked glycans are transformed from high-mannose sugars into a large, diverse repertoire of hybrid and complex N-linked glycan types that are then secreted and presented on the cell surface as glycosylated proteins as illustrated in Figure 1-7 below.^{49, 51}

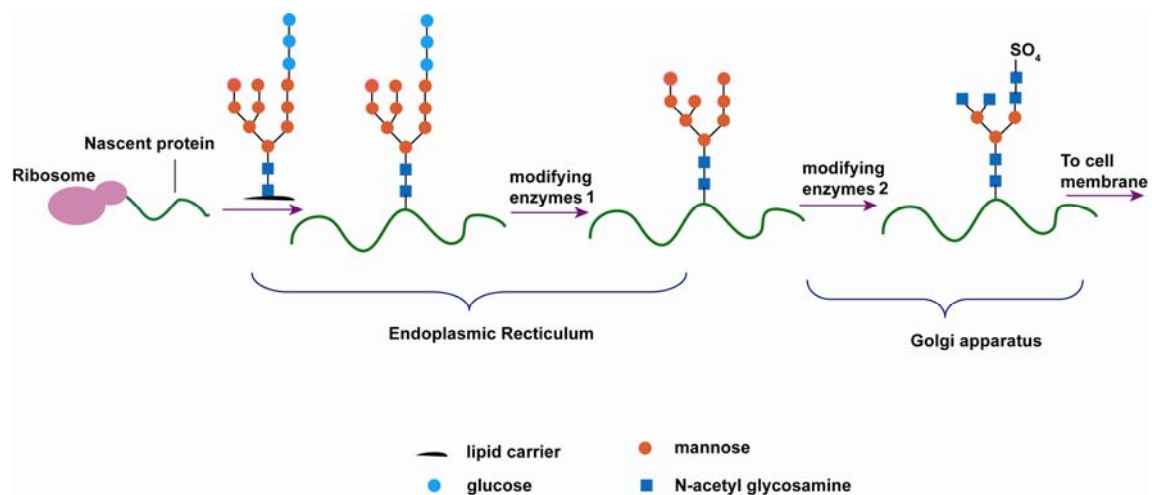


Figure 1-7: Biosynthesis of N-glycosylated proteins. The nascent protein is glycosylated as it grows and this process starts in the ER and ends in the GA after which the glycosylated protein is presented on the cell surface. The structures shown after modifying enzymes arrows represent one of the many possible examples of what the glycans could be.

As a result, the secreted glycosylated protein contains a diverse population of glycoforms that may consist of one to several different types of N-linked glycans attached at the same or different glycosylation site(s). The end result is a glycosylated protein that is structurally complicated hence very challenging to characterize. This is one of the main reasons why the glycoproteomics field has been lagging behind the field of proteomics.

As shown in Figure 1-8, all N-linked glycans have a common pentasaccharide (tri-antennary) core and only vary based on the type and number of sugars added and how they are branched from the tri-antennary core. Depending on substitution of the non-reducing terminal residues in this core structure, the N-linked glycans can be classified as complex, high-mannose, and hybrid type as shown in Figure 1-8. The high-mannose type contain the most number of mannose residues, while the hybrid and complex types are often processed further by addition of terminal residues like sialic acid or sulfate, and sometimes contain a fucose linked to GlcNAc-Asn.

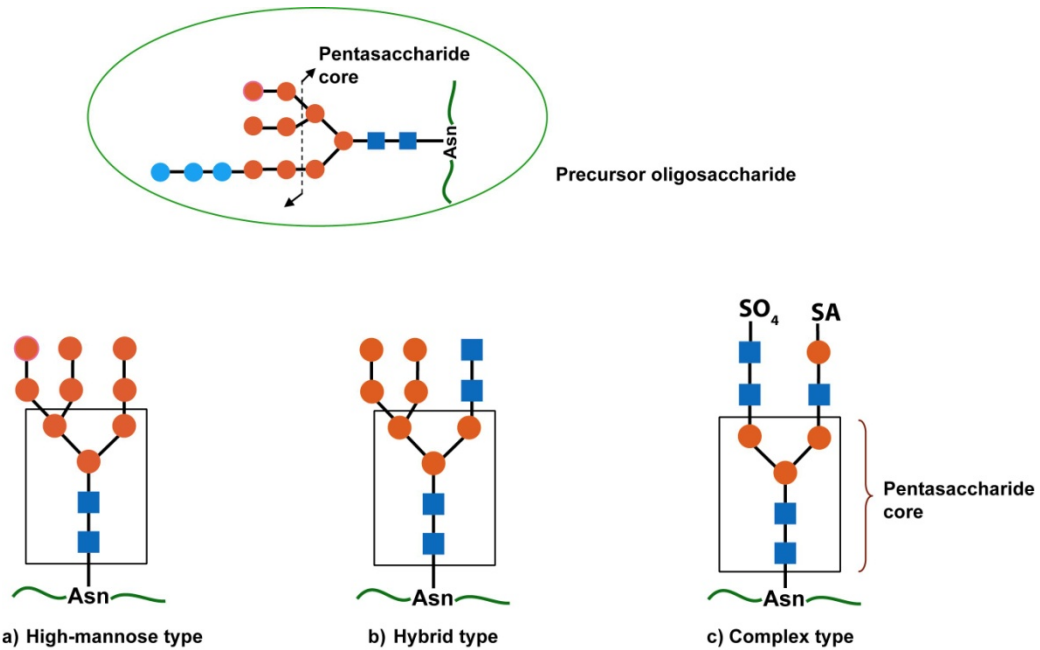


Figure 1-8: Types of N-linked glycans. All N-linked glycans start from the precursor oligosaccharide which is then trimmed down to a) high-mannose, b) Hybrid, and, c) Complex type. The broken arrow indicates the building block for all N-linked glycans. (also boxed pentasaccharide core)

1.5.2 Factors affecting biosynthesis of N-glycosylated proteins

The biosynthesis of N-glycosylation in proteins is a highly specific process that depends on the specific cell type expressing the protein, the specific polypeptide chain of the protein to be glycosylated (which usually contains encoded information that directs its glycosylation) and the specific glycosylation site to be glycosylated.^{49, 51} Other factors that affect protein N-glycosylation include protein conformation which may affect the accessibility of potential glycosylation sites to modifying enzymes, transport rates in endoplasmic reticulum (ER) and Golgi, cellular regulators that affect glycosyl-transferases' activity, and localization of glycosyltransferases

(availability of modifying enzymes and the order in which they interact with the protein).^{49, 51} All these factors also contribute to structural diversification of N-linked glycans.

1.5.3 Biological Importance of N-linked glycosylated proteins

In humans, approximately 50-60% of all the proteins are glycosylated.^{52, 53} The glycan content in these proteins accounts for about 4-60% of the total protein mass. The presence of these glycans in proteins is known to have profound influence on the physiochemical, cellular and biological functional properties of proteins. For instance, glycans play key roles in physiochemical properties of proteins such as stability, folding, conformation, solubility, and protecting proteins against proteolysis.⁵⁴ In addition, glycans also influence biological processes in proteins like facilitating cell-cell recognition, recognition of hormones, toxins, viruses, coordination of immune functions; glycans are also involved embryonic protein-glycan interaction.⁵⁵⁻⁵⁸ In addition, variations in glycan structures have been observed in several pathological states such as cancers, rheumatoid arthritis, carbohydrate-deficient glycoprotein syndromes (CDGS).⁴⁹ So far, there are several glycoproteins that have or are being investigated in biomedical research for disease prognosis and for therapeutic purposes. For instance, about 25% of the currently approved cancer biomarkers are glycosylated proteins.⁵⁹ Consequently, defining the

structures and locations of these glycans in proteins may provide useful insights into how variation in glycosylation affects the functions of proteins in health and disease thus provide valuable information that can be useful in facilitating the understanding of structure-function relationship. In order to acquire this information, sensitive, rapid and reliable methods for mapping and profiling glycosylation in proteins are required.

1.6 An overview and summary of the following chapters

The work presented herein focuses on developing mass spectrometric methods to characterize glycans in different glycoproteins in a glycosylation site-specific fashion. This approach allows structural elucidation of both the glycan and their attachment site in a single-MS experiment. Characterizing the glycans in glycoproteins in a glycosylation site-specific manner is important because the functional role of the similar glycans structures varies from protein to protein as there is no general function that can be attributed to similar glycans in different proteins or different glycans in the same protein.⁶⁰ In order to understand the structure-function relationship of glycans in glycoproteins, structural analysis is important. As a result, it is critically important to not only characterize glycans but also determine their precise locations on each protein. The studies described herein, mainly focused on two important classes of glycoproteins: glycoproteins hormones and glycoproteins found on the surface of a virus such as HIV envelope proteins.

The main interest in characterizing glycans in glycoprotein hormones and HIV envelope proteins stems from their biological significance.

The glycoprotein hormones analyzed here are from a family of heterodimeric glycoproteins that consist a non-covalently linked alpha and beta subunit.⁶¹ They consist of luteinizing hormones (LH), thyroid stimulating hormone (TSH), follicle stimulating hormone (FSH), and chorionic gonadotropin (CG). LH, FSH, and TSH are all secreted from the anterior pituitary gland whereas CG is secreted from the placenta.⁶² The pituitary hormones are involved in regulating reproductive and metabolic functions in the body.^{58, 61, 63} Structural studies on these hormones indicate that glycans account for 15-35% of the total molecular weight of these hormones. These glycans contain unusually high content of terminal acidic residues such as sulfate groups and sialic acid.⁶⁴ For example, TSH contains glycans that are exclusively sulfated while the FSH contains glycans that are both sialylated and sulfated. These terminal residues act as unique features on these glycans that are recognized by specific receptors. For instance, glycans capped with a sulfate group are recognized by a receptor in the liver, facilitating their plasma clearance.⁶⁵ Therefore, the addition of a sulfate group targets these hormones for rapid removal from the body. In addition, the presence of these terminal residues have been associated with several diseases such as cancer, rheumatoid arthritis etc.⁴⁹ However, the precise differences in degree of sulfation or sialylation between healthy and infected

individuals and the role that these residues play are still not well understood. This is mainly due to the acidity and lability of these groups, creating a significant analytical challenge that has greatly hindered the analysis of these species. We have developed and validated MS methodologies for characterizing glycans containing these terminal residues in a glycosylation site-specific manner and successfully applied them in characterizing glycoprotein hormones.⁶⁶ These MS methods are not only applicable to glycoprotein hormones but can also be applied to any glycoprotein containing these terminal residues.

The other class of glycoproteins that we have characterized are glycoproteins found on the surface of HIV virus. HIV virus is a retrovirus that causes acquired immune deficiency syndrome (AIDS), which has led to a world epidemic that is claiming millions of human lives every year.^{67, 68} The entry and fusion of this virus to its target host cells is mediated by an envelope glycoprotein, gp160, which consists of an exterior envelope protein (gp120), anchored onto the viral membrane by a trans-membrane envelope protein (gp41).⁶⁹⁻⁷¹ The exterior surface envelope, gp120, is exposed to the immune system of the host cells and is therefore the main focus for HIV vaccine development.^{68, 72} This envelope protein is one of the most glycosylated proteins known in nature, with over 50% of its mass consisting of glycans.^{73, 74} It has been suggested that the high population and diverse range of glycan structures on this protein act as a shield for the virus and

protect it from proteolytic degradation.^{74, 75} In addition, these glycans act as the major defense mechanism for the virus by masking the underlying epitopes on the protein backbone making them invisible to the immune system.⁷⁵⁻⁸¹ The glycan shield is also said to evolve during infection and throughout the HIV disease progression, and this evolution can result in changes in glycan structures, position, and/or number of glycosylation sites.^{78, 81-86} As a result, mapping the glycosylation sites on HIV envelope proteins, determining the structures, populations, and location of these glycans on the protein is important and may yield valuable information that may contribute in developing the long anticipated HIV vaccine. Unfortunately glycoprotein analysis is generally a very challenging task^{44, 87, 88} and among all the N-linked glycoproteins that have been studied so far, gp120 envelope protein is by far the most complicated. This is due to its protein sequence variation and the high number of potential glycosylation sites (over 24 potential glycosylation sites), each containing a high population of a wide range of glycan structures that can evolve with time,^{73, 81, 89-94} therefore posing a great challenge to most analytical methods that are currently available. We have developed different MS techniques to characterize N-glycans of one of the potential candidates for an HIV vaccine, CON-S gp140 Δ CFI, which is derived from gp160, in a glycosylation site-specific manner. By employing these strategies, we have successfully characterized hundreds of N-glycans, and detected all the 31 potential glycosylation sites

present in this protein. The glycan profiles of this vaccine candidate can be correlated to immunogenicity, as a first step in using glycosylation information to aid vaccine development.

Chapter 2

A new approach to characterize sulfated glycans present on glycoproteins is described. The analysis is performed on glycopeptides, so information about the sulfated species was obtained in a glycosylation site-specific manner. Typically, negative ion mode is the method of choice for analysis of sulfated glycans since negatively charged species ionize more efficiently in the negative ion mode. However, (-) MS/MS for sulfated glycopeptides provides limited structural information due to the lability of the SO_3 group. To overcome this problem, a method that employs an ion-pairing reagent to stabilize the SO_3 group of the glycopeptides was developed thereby promoting other dissociation pathways that provide more structural information. The amount of structural information obtained from (+) ESI-MS/MS of the ion-pair complexes for sulfated glycopeptides of equine thyroid stimulating hormone (eTSH) was compared with information obtained by (-) ESI-MS/MS of the un-derivatized, sulfated glycopeptides. The results indicated that this new method provides detailed insights into the sequence, branching and type of N-glycans present, compared to analysis via (-) ESI-MS/MS.⁶⁶

Chapter 3

Mass spectral analysis is an increasingly common method used to characterize glycoproteins. When more than one glycosylation site is present on a protein, obtaining MS data of glycopeptides is a highly effective way of obtaining glycosylation information, because this approach can be used not only to identify *what* the carbohydrates are, but also *at which glycosylation site* they are attached. Unfortunately, this is not yet a routine analytical approach, in part because data analysis can be quite challenging. We have developed strategies to simplify this analysis. Presented herein is a novel mass spectrometry technique that identifies the peptide moiety of either sulfated, sialylated or both sialylated and sulfated glycopeptides. This technique correlates product ions in collision induced dissociation (CID) experiments of suspected glycopeptides to a peptide composition, using a newly developed web-based tool, GlycoPep ID. After identifying the peptide portion of glycopeptides with GlycoPep ID, the process of assigning the rest of the glycopeptide composition to the MS data is greatly facilitated because the “unknown” portion of the mass assignment that remains can be directly attributed to the carbohydrate component. Several examples of the utility and reliability of this method are presented herein.⁹⁵

Chapter 4

CON-S gp140 Δ CFI is a potential candidate for HIV/AIDS vaccine and to our knowledge, it is the most N-glycosylated protein characterized so far, with 31 potential glycosylation sites. While the protein sequence of this protein is well known, the glycans shielding its surface, which accounts for about 50% of its molecular weight, have not yet been characterized. Furthermore, although mass spectrometry has gained a widespread use in glycoprotein analysis, so far there is no consensus as to which mass spectrometry method gives the best glycosylation coverage. As a result, to characterize this important protein, the two most widely used MS techniques, LC/ESI-MS and MALDI-MS were employed and the results obtained therewith compared to determine which of the two techniques would be more suitable for analysis of a complicated glycoprotein like CON-S gp140 Δ CFI in providing the most glycosylation information content in this protein in terms of sequence coverage, number and type of glycans.

1.7 References

1. Siuzdak, G. The expanding the role of mass spectrometry in biotechnology. San Diego: MCC 2003.
2. De Hoffmann, E.S. Mass spectrometry: Principles and applications. Second Ed. New York: John Wiley & Sons, 2002
3. Glish, G.L; Vachet, R.W. *Nat Rev Drug Discov* **2003**, 2, 140-150
4. Dass, C. Fundamentals of contemporary mass spectrometry; Hoboken, New Jersey, John Wiley & Sons, Inc: 2007.

5. Fenn, J. B; Mann, M; Meng, C.K; Wong, S.F; Whitehouse, C.M. *Science* **1989**, 246, 64-71
6. Henry, K.D; Williams, E.R; Wang, B.H; McLafferty, F.W; Shabanowitz, J; Hunt, D.F. *Proc Natl Sci Acad U S A* **1989**, 86, 9075-9078
7. Loo, J.A; Edmonds, C.G; Smith, R.D. *Anal Chem* **1991**, 63, 2488-2499
8. Blakley, C.R; Carmody, J.J; Vestal, M.L. *J Am Chem Soc* **1980**, 102, 5931-5933
9. Edmonds, C.G; Loo, J.A; Loo, R.O; Udseth, H.R; Barinaga, C.J; Smith, R.D. *Biochem Soc Trans* **1991**, 19, 943-947
10. Cole, R.B. *Electrospray ionization mass spectrometry: Fundamentals instrumentation and application*. New York: John Wiley & Sons, Inc 1997.
11. Cech, N.B; Enke, C.G. *Mass Spectrom Rev* **2002**, 20, 362-387
12. Ikonomou, M.G; Blades, A.T; Kebarle, P. *Anal Chem* **1991**, 63, 1989-1998
13. Enke, C.G. *Anal Chem* **1997**, 69, 4885-4893
14. Van Berkel, G.J; Zhou, F; Aronson, J.T. *Int J Mass Spectrom Ion Process* **1997**, 162, 55-67
15. Tafllin, D.C; Ward, T.L; Davis, E.J. *Langmuir* **1989**, 5, 376-384
16. Solouki, T; Reinhold, B.B; Costello, C.E; O'Malley, M; Guan, S; Marshall, A.G. *Anal Chem* **1998**, 70, 857-864
17. McLafferty FW. *Int J Mass Spectrom* **2001**, 212, 81-87
18. Easterling, M.L; Amster, I.J. *NATO ASI Ser C: Math Phys Sci* **1997**, 504, 287-314
19. Heeren, R.M.A; Kleinnijenhuis, A.J; McDonnell, L.A; Mize, T.H. *Anal Bioanal Chem* **2004**, 378, 1048-1058
20. Comisarow, M.B; Marshall, A.G. *Chem Phys Lett* **1974**, 25, 282-283

21. Amster, I.J. *J Mass Spectrom* **1996**, 31, 1325-1337
22. Marshall, A.G. *AIP Conf Proc* **1998**, 430, 3-13
23. Hendrickson, C.L.; Drader, J.J, Laude, D.A; Guan,S; Marshall, A.G. *Rapid Comm Mass Spectrom* **1996**, 10, 1829-1832
24. Jacoby, C.B; Holliman, C.L; Gross, M.L. *NATO ASI Ser C: Math Phys Sci* **1992**, 353, 93-16
25. Holliman, C.L; Rempel, D.L; Gross, M.L. *Mass Spectrom Rev.* **1994**, 13, 105-132
26. Mclver, R.T.; Baykut, G; Hunter, R.L. *Int J Mass Spectrom Ion Process* **1989**, 89, 343-358
27. Marshall, A.G. *Acc Chem Res* **1985**, 18, 316-322
28. Marshall, A.G; Hendrickson, C.L; Jackson, G.S. *Mass Spectrom Rev* **1998**, 17, 1-35
29. Marshall, A.G; Grosshans, P.B. *Anal Chem* **1991**, 63, 215A-229A
30. Marshall, A.G; Hendrickson, C.L. *Int J Mass Spectrom* **2002**, 215, 59-75
31. Alber, G.M; Marshall, A.G; Hill, N.C; Schwelkhard, L; Ricca, T.L. *Rev Sci Instrum* **1993**, 64, 1845-1852
32. Lebrilla, C.B; Amster, I.J; Mclver, R.T. *Int J Mass Spectrom Ion Process* **1989**, 87, R7-R13
33. Henry, K.D; McLafferty, F.W. *Org Mass Spectrom* **1990**, 25, 490-492
34. Payne, A.H; Glish, G.L. *Method Enzymol* **2005**, 402, 109-148
35. Dey, M; Castoro, J.A; Wilkins, C.L. *Anal Chem* **1995**, 67, 1575-1579
36. Beu, S.C; Laude, D.A. *Anal Chem* **1991**, 63, 2200-2203
37. Laskin, J; Futrell, J.H. *Mass Spectrom Rev* **2005**, 24, 135-167
38. Cody, R.B; Burnier, R.C; Cassady, C.J; Freiser, B.S. *Anal Chem* **1982**, 54, 2225-2228

39. Little, D.P; Speir, J.P; Senko, M.W; O'Connor, P.B; McLafferty, F.W. *Anal Chem* **1994**, 66, 2809-2815
40. Marshall, A.G; Hendrickson, C.L; Emmett, M.R. *Adv Mass Spectrom* **1998**, 14, Chapter 10/221-Chapter 210/239
41. Dunbar, R.C; McMahon, T.B. *Science* **1998**, 279, 194-197
42. Zubarev, R.A. Roman, A; Kelleher, N.L; McLafferty, F.W. *J Am Chem Soc.* **1998**, 120, 3265-3266
43. Hillenkamp, F; Karas, M; Beavis, R.C; Chait, B.T. *Anal Chem* **1991**, 63, 1193A-1203A
44. Budnik, B.A; Lee, R.S; Steen, A.J. *Biochim Biophys Acta*, **2006**, 1764, 1870-1880
45. Morelle, W; Canis, K; Chirat, F; Faid, V; Michalski, J-C. *Proteomics* **2006**, 6, 3993-4015
46. Burlingame, A.L. *Curr Opin Biotechnol* **1996**, 7, 4-10
47. Sullivan, B; Addona; T.A; Carr, S.A. *Anal Chem* **2004**, 76, 3112-3118
48. Zaia, J. *Mass Spectrom Rev* **2004**, 23, 161-227
49. Varki, A.C; R.; Esko, J.; Freeze, H.; Hart, G.; Marth, J. *Essentials of Glycobiology*. New York: Cold Springs Harbor 1999
50. Welply, J.K; Shenbagamurthi, P; Lennarz, W.J; Naider, F. *J Biol Chem* **1983**, 258, 11856-11863
51. Conradt, H.S. *Protein Glycosylation: Cellular, Biotechnological and Analytical Aspects*. New York, VCH: Cambridge 1991
52. Apweiler, R; Hermjakob, H; Sharon, N. *Biochim Biophys Acta* **1999**, 1473, 4-8
53. Hagglund, P; Bunkenborg, J; Elortza, F; Jensen Ole, N; Roepstorff, P. *J Proteome Res* **2004**, 3, 556-566
54. Morelle, W; Michalski, J.C. *Curr Anal Chem* **2005**, 1, 29-57

55. Guo, Y; Feinberg, H; Conroy, E; Mitchell, D.A; Alvarez, R; Blixt, O; Taylor, M.E. *Nat Struct Mol Biol* **2004**, 11, 591-598
56. Collins, B.E; Paulson, J.C. *Curr Opin Chem Biol* **2004**, 8, 617-625
57. Rudd, P.M; Wormald, M.R; Dwek, R.A. *Trends Biotechnol* **2004**, 22, 524-530
58. Baenziger, J.U. *Am J Pathol* **1985**, 121, 382-391
59. Ludwig, J.A; Weinstein, J.N. *Nat Rev Cancer* **2005**, 5, 845-856
60. Howard, S.C; Wittwer, A.J; Welply, J.K. *Glycobiology* **1991**, 1, 411-418
61. Pierce, J.G; Parsons, T.F. *Annu Rev Biochem* **1981**, 50, 465-495
62. Baenziger, J.U; Green, E.D. *Biochim Biophys Acta*, **1988**, 947, 287-306
63. Green, E.D; Baenziger, J.U; Boime, I. *J Biol Chem* **1985**, 260, 15631-15638
64. Green, E.D; Baenziger, J.U. *J Biol Chem* **1987**, 262, 12018-12029
65. Fiete, D; Srivastava, V; Hindsgaul, O; Baenziger, J.U. *Cell* **1991**, 67, 1103-1110
66. Irungu, J; Dalpathado, D.S; Go, E.P; Jiang, H; Ha, H-V; Bousfield, G.R; Desaire, H. *Anal Chem* **2006**, 78, 1181-1190
67. Gallo, R.C; Salahuddin, S.Z; Popovic, M; Shearer, G.M; Kaplan, M; Haynes, B.F; Palker, T.J. *Science* **1984**, 224, 500-503
68. Graham, B.S; Wright, P.F. *N Engl J Med* **1995**, 333, 1331-1339
69. Kwong, P.D; Wyatt, R; Robinson, J; Sweet, R.W; Sodroski, J; Hendrickson, W.A. *Nature* **1998**, 393, 648-659
70. Chan, D.C; Kim, P.S. *Cell* **1998**, 93, 681-684
71. Wyatt, R; Sodroski, J. *Science* **1998**, 280, 1884-1888
72. Pantophlet, R; Burton, D. R. *Annu Rev Immunol* **2006**, 24, 739-769

73. Geyer, H; Holschbach, C; Hunsmann, G; Schneider, J. *J Biol Chem* **1988**, 263, 11760-11767
74. Koch, M; Pancera, M; Kwong, P.D; Kolchinsky, P; Grundner, C; Wang, L; Hendrickson, W.A. *Virology* **2003**, 313, 387-400
75. Kwong, P.D; Doyle, M.L; Casper, D.J; Cicala ,C; Leavitt, S.A; Majeed, S; Steenbeke, T.D. *Nature* **2002**, 420, 678-682
76. Wyatt, R; Kwong; P.D; Desjardins, E; Sweet, R.G; Robinson, J; Hendrickson, W.A; Sodroski, J.G. *Nature* **1998**, 393, 705-711
77. Reitter, J.N; Means, R.E; Desrosiers, R.C. *Nat Med* **1998**, 4, 679-684
78. Wei, X; Decker, J.M; Wang, S; Hui, H; Kappes, J.C; Wu, X; Salazar-Gonzalez, J.F. *Nature* **2003**, 422, 307-312
79. Stamatatos, L; Wiskerchen, M; Cheng-Mayer, C. *AIDS Res Hum Retroviruses* **1998**, 14, 1129-1139
80. Back, N.K; Smit, L; De Jong, J.J; Keulen, W; Schutten, M; Goudsmit, J; Tersmette, M. *Virology* **1994**, 199, 431-438
81. Cheng-Mayer, C; Brown, A; Harouse, J; Luciw, P.A; Mayer, A.J. *J Virol* **1999**, 73, 5294-5300
82. Robinson, W.E; Montefiori, D.C; Mitchell, W.M. *AIDS Res Hum Retroviruses* **1987**, 3, 265-282
83. Ly, A; Stamatatos, L. *J Virol* **2000**, 74, 6769-6776
84. Overbaugh, J; Rudensey, L.M. *J Virol* **1992**, 66, 5937-5948
85. Overbaugh, J; Rudensey, L.M; Papenhausen, M.D; Benveniste, R.E; Morton, W.R. *J Virol* **1991**, 65, 7025-7031
86. Sagar, M; Wu, X; Lee, S; Overbaugh, J. *J Virol* **2006**, 80, 9586-9598
87. Alvarez-Manilla, G; Atwood, J; Guo, Y.; Warren, N.L; Orlando, R; Pierce, M. *J Proteome Res* **2006**, 5, 701-708
88. Morelle, W; Michalski, J.C. *Curr Pharm Des* **2005**, 11, 2615-2645

89. Mizuochi, T; Matthews, T.J; Kato, M; Hamako, J; Titani, K; Solomon, J. Feizi, T. *J Biol Chem* **1990**, 265, 8519-8524
90. Leonard, C.K; Spellman, M.W; Riddle, L; Harris, R.J; Thomas, J.N; Gregory, T.J. *J Biol Chem* **1990**, 265, 10373-10382
91. Zhu, X; Borchers, C; Bienstock, R.J; Tomer, K.B. *Biochemistry* **2000**, 39, 11194-11204
92. Yeh, J-C; Seals, J.R; Murphy, C.I; van Halbeek; H, Cummings, R.D. *Biochemistry* **1993**, 32, 11087-11099
93. Cutalo, J.M; Deterding, L.J; Tomer, K.B. *J Am Soc Mass Spectrom* **2004**, 15, 1545-1555
94. Gao, F; Weaver, E.A; Lu, Z; Li, Y; Liao, H-X; Ma, B; Alam, S M. *J Virol* **2005**, 79, 1154-1163
95. Irungu, J; Go, E. P; Dalpathado, D. S; Desaire, H. *Anal Chem* **2007**, 79, 3065-3074

CHAPTER 2

A Method for Characterizing Sulfated Glycoproteins in a Glycosylation Site-Specific Fashion, Using Ion-Pairing and Tandem Mass Spectrometry

Reprinted by permission from *Anal Chem* **2006**, 78, 1181-1190. Copyright 2006 American Chemical Society.

2.1 Introduction

Glycoproteins bearing sulfated carbohydrates have been identified from different species, ranging from bacteria to humans.¹⁻³ Changes in sulfation have been linked to osteoarthritis,⁴⁻⁶ cystic fibrosis, and cancer.⁷⁻⁹ Glycoprotein hormones, particularly lutropin (LH) and thyrotropin (TSH), are heavily sulfated hormones that regulate reproduction and metabolism.^{1,10,12} These hormones consist of two non-covalently linked α and β subunits, each containing one to two glycosylation sites. The degree of sulfation at each of these sites varies significantly within a given hormone. For example the single glycosylation site at the β subunit in lutropin contains the greatest proportion of disulfated glycans, whereas one of the two sites on the α subunit contains the greatest percentage of monosulfated glycans.¹⁰⁻¹³ The presence of the sulfate groups in these hormones can alter their biological recognition and facilitate their rapid clearance from the body.¹⁴ As a result, studying sulfation at each of these sites is important to understand how the carbohydrate structure affects the function of these hormones. To date, an

analytical method is not available that can both identify sulfation in a glycosylation site-specific fashion, and characterize the sulfated glycans. Therefore, developing efficient and sensitive analytical techniques that are capable of identifying and characterizing sulfated species in a glycosylation site-specific manner are needed, in order to facilitate the understanding of their biological significance.

One of the most common biochemical methods to identify sulfated glycans is metabolic radiolabeling, followed by fractionation and characterization of the resulting glycans.^{9-11,13,15-18} This method has the advantage of distinguishing the presence of isomeric structures, which can be very useful for structural elucidation. Although highly effective, this approach can be hazardous and time consuming. Most often, glycans are released from the protein using chemical or enzymatic procedures. As a result, information is not available about which glycans originated from which glycosylation sites, unless the glycosylation sites are separated before glycan release.

Mass spectrometry (MS) has emerged as an important analytical tool in the structural elucidation of glycoconjugates. It has the advantage of high sensitivity, speed, and low sample requirements.¹⁹ Typically, MS analysis of glycoproteins is done after the glycans are released from the polypeptide backbone.²⁰ This technique has been previously applied to analyze sulfated glycans using fast atom bombardment (FAB),²¹ electrospray ionization

(ESI),³ and matrix assisted laser desorption-ionization (MALDI).²² While this method is highly effective, it fails to provide glycosylation site-specific information that is essential in understanding structure/function relationship of glycans in glycoproteins.

An alternative approach, which has the advantage of providing site specific information, involves proteolytic digestion of the glycoprotein followed by MS analysis of the resulting glycopeptides.^{20,23-26} Unfortunately the applicability of this strategy in the analysis of sulfated glycopeptides is limited because proteolysis liberates both sulfated and non-sulfated glycoforms, and the signal of the sulfated glycopeptides is suppressed in both positive and negative ion mode during MS analysis.²⁴ To overcome this limitation, an MS strategy that employs an ion-pairing reagent to enhance the signal of sulfated glycoforms has been recently developed.²⁴

In addition to signal suppression, another challenge that has lead to the limited applicability of direct mass spectrometric analysis of sulfated glycopeptides is the fact that MS/MS data of sulfated glycopeptides are very different than MS/MS data acquired from related species, including nonsulfated glycopeptides *and* sulfated glycans.²⁴ It is well known that sulfated species ionize better in the negative ion mode; however, preliminary studies of (-)ESI-MS/MS spectra of sulfated glycopeptides indicate that very minimal structural information is obtained.²⁴ This may be attributed to the fact that dissociation pathways in the negative ion mode are driven by the

deprotonation site. Since the site of deprotonation is not on the carbohydrate backbone, glycosidic cleavages, which would provide structural information about the carbohydrate, are not readily observed.²⁴ Thus, structural characterization of sulfated glycopeptides using traditional methods is not a straightforward task.

Herein we demonstrate that by using an ion-pairing reagent, sulfated glycopeptides can be analyzed in positive ion mode, and significant structural information is obtained during MS/MS analysis. While ion-pairing has been used previously to enhance the MS signal of sulfated compounds^{24, 27-30}, and to discriminate sulfation from phosphorylation,³¹ this is the first report that demonstrates MS/MS of the ion-pair complexes provides structural information for the complexed analytes. In this report, sulfated glycopeptides from horse TSH, possessing glycan structures identical to those characterized in bovine and human TSH,^{11,22} were subjected to MS/MS analysis in both negative ion mode (without ion-pairing reagent) and in positive ion mode, after the addition of the ion-pairing reagent. The information obtained from the two MS/MS techniques is described. To demonstrate the general applicability of these studies, the fragmentation trends that are described for (-) ESI-MS/MS and ion-pairing MS/MS were used to characterize the structures of two unknown, sulfated glycopeptides.

2.2 Experimental

2.2.1 Digestion of eTSH with Proteinase K

Glycopeptides from equine thyroid stimulating hormone (eTSH) were generated in the laboratory of Dr. George Bousfield, Wichita State University. Briefly, the eTSH glycoprotein was reduced and alkylated based on a method described previously.³² The glycoprotein was desalted using centrifugal ultrafiltration, digested with proteinase K (10%w/w) and dried.²⁵ The eTSH digests were then subjected to Superdex peptide gel filtration chromatography. The fraction containing carbohydrates was collected, dried, and analyzed as described below.

2.2.2 Peptide Sequencing (Edman)

The peptide sequences of these glycopeptides were verified by automated Edman degradation using an Applied Biosystems (Foster City, CA) model 492 Procise sequencer. The eTSH digests were applied to Biobrene-coated glass fiber membranes that had been precycled in the sequencer. Typical sequencer experiments consisted of seven automated Edman degradation cycles, sufficient to sequence the entire length of 3-5 residue peptides resulting from proteinase K digestion.

2.2.3 eTSH glycopeptide preparation for MS analysis

The dried glycopeptides were first dissolved in water and diluted with MeOH:H₂O (4:1) containing 0.3% acetic acid, to constitute a final concentration of 0.03 µg/µL. This solution was used directly for (-) ESI-MS

analysis. It was also used in the ion-pairing experiments. For ion pairing, a tripeptide (Lys-Lys-Lys or 3K), purchased from Sigma-Aldrich (St. Louis, MO), was first dissolved in H₂O and diluted to 0.1 µg/µL MeOH:H₂O (4:1) containing 0.3% acetic acid. Ion-pair complexes were formed by combining equal volumes of the basic peptide and the glycopeptide solutions. The mixture was vortexed prior to injecting into the mass spectrometer.

2.2.4 Mass Spectrometry

MS data was acquired on a high-resolution Thermo Finnigan linear ion trap-Fourier Transform Ion Cyclotron Resonance mass spectrometer, LTQ-FTICR, (San Jose, CA) equipped with a 7 Telsa actively shielded magnet. Samples were directly infused into the mass spectrometer using a syringe pump at a flow rate of 5 µL/min. High resolution data was acquired by maintaining resolution at 50,000, for m/z 400. The instrument was externally calibrated prior to the analysis, over the entire mass range of interest. The data was acquired in the mass range of m/z 800-2000. Electrospray ionization in negative mode was achieved using a spray voltage of approximately -4.0 kV. N₂ was used as a nebulizing gas at 20 psi, and the capillary temperature was maintained between 200-230 °C. Data was acquired and processed using Xcalibur 1.4 SR1 software (Thermo Finnigan San Jose, CA). The glycopeptide compositions were assigned for the peaks in the high resolution data by using a visual basic algorithm, developed previously.²⁵

Data was also acquired in the positive ion mode after the addition of an ion-pairing reagent. Electrospray ionization in positive mode was achieved using a spray voltage of 4.0 kV. All ions that corresponded to ion-pair complexes of the ions observed in (-) ESI-MS were subjected to (+) ESI-MS/MS.

2.2.5 MS/MS analysis

MS/MS experiments were performed using the linear ion trap mass analyzer, on the LTQ-FTICR MS, to confirm the assigned glycopeptide compositions and to obtain structural information. The MS/MS data was acquired in both negative mode (without ion-pairing reagent) and in positive ion mode, after the ion-pairing reagent was added. In both cases, the precursor ions were activated for 30 ms with activation Q of 0.25 and an isolation width of 3 Da. Activation amplitudes were in the range of 19-29 % as defined by the instrument software.

2.3 Results and Discussion

2.3.1 Compositional analysis of eTSH glycopeptides

Edman (N-terminal) chemistry provides reliable sequence information for amino acids in an unknown peptide.³³ This technique was used to verify the actual sequence of the amino acids of the peptide moieties attached to the glycan structures generated in this study. Three main peptide sequences consisting of LENHTQ, NIT and TINTT were identified, which corresponded

to the glycosylation sites at α Asn⁸², α Asn⁵⁶, and β Asn²³, respectively. Based on the Edman sequencing data, it was apparent that small quantities of several other peptide sequences were also present. These corresponded to shorter variants of these glycopeptides.

The peptide sequences obtained from Edman sequencing data were exported to a visual basic algorithm, which was used in conjunction with the high resolution FTICR-MS data (See Figure 2-1) to obtain all the reasonable glycan compositions that matched previously characterized glycan structures in the bovine and human hormones. The identified glycopeptide structures are shown in Table 2-1. The calculated mass error for all the assigned compositions was less than 2 ppm. The data in this table indicates that a heterogeneous mixture of the glycopeptides, all of which were either mono- or di-sulfated, were identified from this hormone.

Figure 2-1 shows the mass spectrum acquired in the negative ion mode of the identified glycopeptides found in eTSH. Glycans from two glycosylation sites were identified corresponding to N⁵⁶IT and TIN²³TT of the α and β subunit of this hormone, respectively, and they are uniquely labeled on the mass spectrum. The majority of the glycan structures attached to the α subunit peptide, N⁵⁶IT, were hybrid monosulfated glycans, whereas the one identified from the β subunit (TIN²³TT) was a disulfated, fucosylated, complex-type glycan. No glycans were identified corresponding to the LEN⁸²HTQ site.

The glycans' structures from the hormone (TSH) used in this study have been previously characterized by Green and Baenziger.¹¹ In the previous analysis, the metabolically radiolabeled sulfated glycans were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and their structures were elucidated by endo- and exo-glycosidase digestion in combination with high performance liquid chromatography (HPLC).¹¹ The same glycan structures were obtained by Harvey et. al using MALDI-MS.²³ Because each of the glycan's structures from this hormone have been fully characterized previously, matching the masses of the glycan portions to the already identified glycan structures allows us to fully know the sequence, branching and the linkage of the monosaccharides within each structure. Therefore, the glycopeptides generated from this hormone can be considered as an ideal set of "standards," for analyzing the fragmentation trends of these compounds, during MS/MS analysis.

These glycopeptides vary with respect to carbohydrate type, number of sulfate groups, and peptide moiety present (see Table 2-1), and the effect of each of these features on the MS/MS data, along with the effect of the charge state, is compared herein. This investigation will enable other investigators to use (-) ESI-MS/MS and ion-pairing MS/MS to characterize the structures of unknown sulfated glycopeptides.

2.3.2 Comparison of (-) ESI-MS/MS data based on

2.3.2.1 Number of SO₃ groups present

Figure 2-2(a) and (b) are (-) ESI-MS/MS data of the doubly charged ions containing either one or two SO₃ groups. The two glycopeptides at *m/z* 942 and 1104 have the same peptide moiety but differ in their glycan composition and the number of SO₃ groups. Had the composition of the singly sulfated glycopeptide been unknown, the MS/MS data would potentially be confusing since the loss of SO₃ is not observed in Figure 2-2(a). Instead, only a few cross-ring cleavages were observed, and the base peak at *m/z* 920 is due to loss of CO₂. Other product ions in the spectrum at *m/z* 1456, 892, and 428 correspond to ^{0,2}A, ^{1,3}A, and ^{0,2}X cross-ring cleavages, respectively. Since ^{0,2}X and ^{0,2}A are complementary ions, resulting from cleavage of the carbohydrate that is attached to the peptide, the presence of either or both ions could be used to identify the peptide moiety of the glycopeptide. The ^{0,2}X ion corresponds to the peptide moiety plus 83 Da (a portion of the carbohydrate) whereas the ^{0,2}A ion is the remaining portion of the carbohydrate. In the case of an unknown N-linked glycopeptide structure, if the mass difference between the precursor and the product ions in (-) MS/MS is greater than 196 (Asn+83), and if this difference does not correspond to a glycosidic cleavage, the observed ion can be assigned as a cross-ring (^{0,2}A) cleavage. The peptide moiety identified for this particular glycopeptide corresponded to NIT. (While only one example of

a singly sulfated, doubly charged glycopeptide is shown, all other similar compounds in this study underwent the same dissociation pathways). Based on these data, it is evident that (-) MS/MS data for monosulfated glycopeptides can only identify the peptide and confirm the presence of a glycopeptide, but it cannot verify whether it is sulfated or not. This is a significant limitation: Detecting the presence of the SO₃ group is essential, because incorporation of this group on the glycan portion of a glycoprotein transforms that glycan into a unique structure that has the potential to be recognized by a specific receptor.¹

The disulfated ion at m/z 1104 in Figure 2-2(b) indicates a loss of SO₃ group from the base peak. This information is useful in confirming that this species is sulfated. Other prominent peaks at m/z 890 (^{0,2}A) and 1926 (Y_{5α}) are due to cross-ring and glycosidic cleavages respectively. The ^{0,2}A ion could be used to identify the peptide moiety, which is NIT for this case, and the Y_{5α} ion is due to the loss of HexNAc and SO₃, indicating that this glycopeptide is sulfated on a terminal HexNAc (assuming this loss was due to one bond cleavage).

These results show that the structural information obtained from the mono- and di-sulfated glycopeptides is dependent on the number of SO₃ groups present. Regardless of the number of sulfates present, the structural information obtained for the glycan portion was minimal, using this method.

2.3.2.2 Different glycosylation sites

(-)ESI-MS/MS data of two disulfated glycopeptides from different glycosylation sites shown in Figure 2-2(b) and (c) indicate that a $^{0,2}A$ ion is observed in both cases. This ion can be used to identify the peptide moieties and therefore distinguish between the two glycosylation sites. While the presence of $Y_{5\alpha}$ and $Y_{5\alpha,5\beta,1\beta}$ ions at m/z 1926 and 1846 (Figure 2-2(b), (c)), respectively, are useful in confirming these ions are glycopeptides, they are inadequate in providing significant structural information of the glycan portions attached to the two identified peptides (NIT and TINTT). Since the amount of structural information obtained from the two glycosylation sites was similar, it was concluded that the fragmentation patterns observed during (-) ESI-MS/MS experiments of these species is independent of the peptide moiety.

2.3.2.3 Different charge states

To maximize the amount of structural information, MS/MS data from all observed charge states were compared. As shown in Figure 2-2(a), no conclusive structural information was obtained from the doubly charged monosulfated ion (m/z 942). However, substantial structural information was obtained from the singly charged ion (m/z 1885) as illustrated in Figure 2-2(d). In addition to the cross-ring cleavages observed from the doubly charged ions, several glycosidic cleavages and a neutral loss of SO_3 were

observed. These ions can be used to verify the composition, identify the partial sequence of the monosaccharides present and determine the type of N-glycan present. For example, the presence of $Y_{5\alpha}$ (m/z 1602) and $Y_{4\beta}$ (m/z 1723), eliminate the possibility that this is a complex-type glycan. Assuming that the trimannosyl chitobiose core is intact, these ions indicate that the type of N-linked glycan must be a hybrid structure. Furthermore, $^{0,2}A$ (m/z 1456) and B_5 (m/z 1336) ions can collectively give information on the peptide moiety present in the glycopeptide (B_5 corresponds to loss of the peptide moiety attached to a core HexNAc). These results indicate that the amount of structural information obtained depends on the charge state of the ion in question, and the singly charged glycopeptides provide more structural information than the doubly charged species. (Other singly charged glycopeptides in this study also produced MS/MS data that could be used to obtain a similar amount of structural information about the glycopeptide).

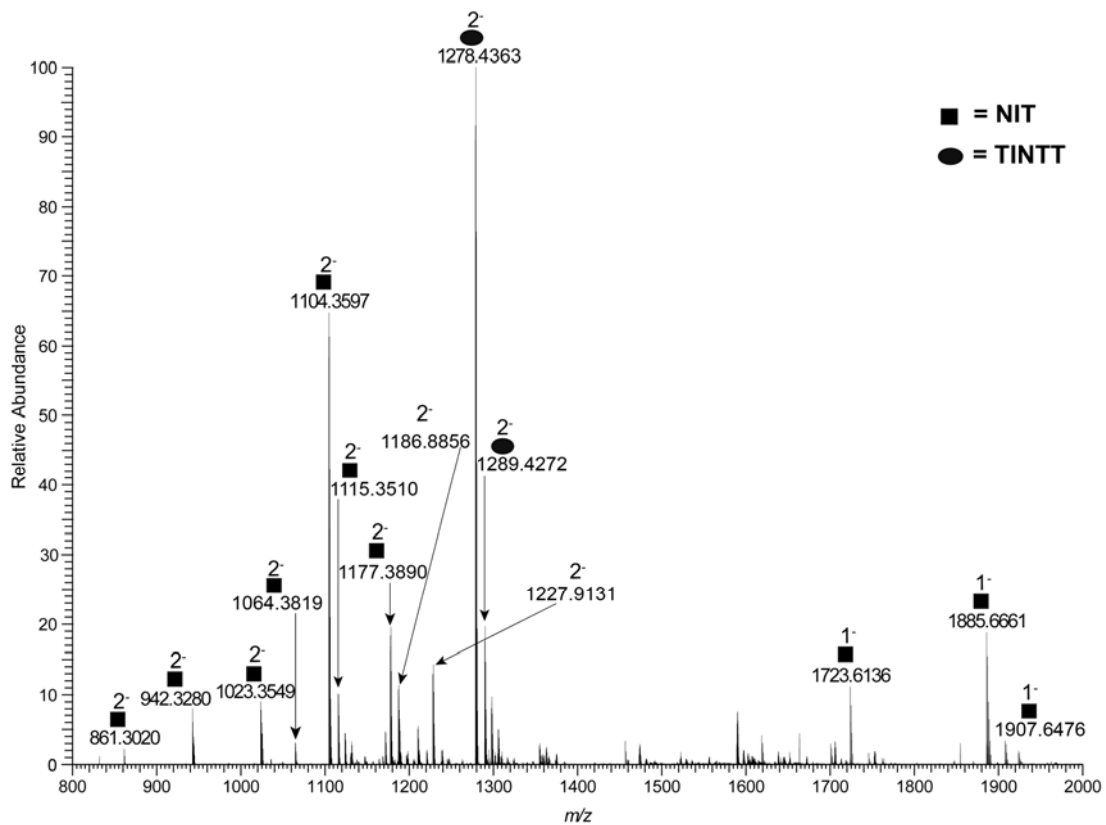


Figure 2-1. FTICR-MS of sulfated glycopeptides from eTSH. The peaks with ovals represent glycans from N⁵⁶IT glycosylation site of the alpha subunit. Peaks with squares represent glycans from TIN²³TT glycosylation site of the beta subunit of eTSH.

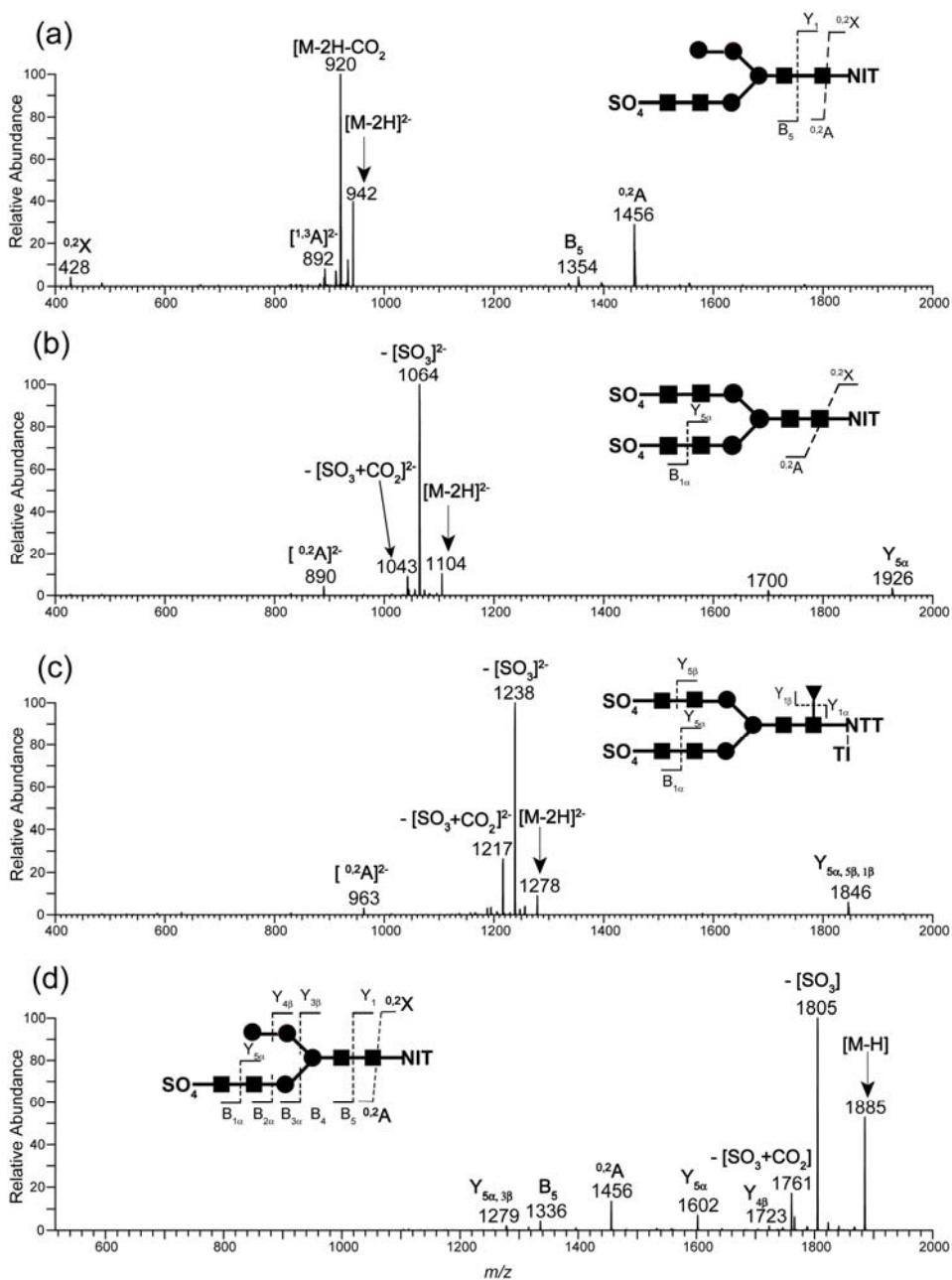
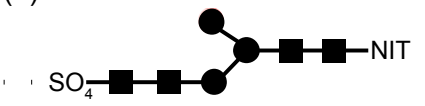
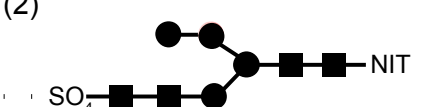
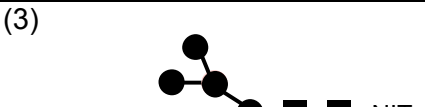
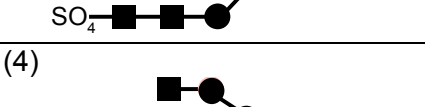
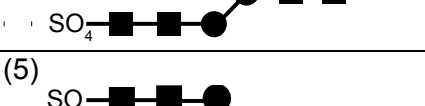
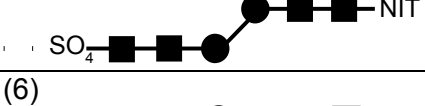
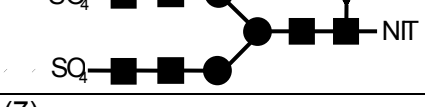
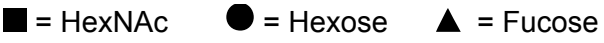


Figure 2-2. Comparison of (-)ESI-MS/MS data for sulfated glycopeptides. (a) MS/MS of a doubly charged monosulfated glycan from N⁵⁶IT glycosylation site; (b) MS/MS of a doubly charged di-sulfated glycan from N⁵⁶IT glycosylation site; (c) MS/MS of a doubly charged disulfated glycan from TIN²³TT glycosylation site; (d) MS/MS of singly charged monosulfated glycan from N⁵⁶IT glycosylation site. Description of the symbols used for the structural formulae can be found in the foot note to Table 2-1.

Table 2-1: Sulfated glycopeptides of eTSH identified using (-)ESI-FTMS

Glycopeptide structures	Charge carrier	Observed m/z	Calculated m/z	Mass error (ppm)
(1) 	[-2H] [-H]	861.3020 1723.6136	861.3012 1723.6102	0.9288 1.9726
(2) 	[-2H] [-H] [+Na ⁺ -2H]	942.3280 1885.6661 1907.6476	942.3272 1885.6630 1907.6449	0.8490 1.6440 1.4154
(3) 	[-2H]	1023.3549	1023.3540	0.8795
(4) 	[-2H]	1064.3819	1064.3805	1.3153
(5) 	[-2H] [+Na ⁺ - 3H]	1104.3597 1115.3510	1104.3589 1115.3499	0.7244 0.9862
(6) 	[-2H]	1177.3890	1177.3879	0.9343
(7) 	[-2H] [+Na ⁺ -3H]	1278.4363 1289.4272	1278.4356 1289.4266	0.5475 0.4653
				

2.3.3 Useful structural information based on fragmentation

characteristics observed in (-)ESI-MS/MS

As described earlier, the presence of $^{0,2}X$ and/or $^{0,2}A$ ions, observed in all cases, can be used as diagnostic ion/s for identifying the peptide moiety of the glycopeptide. This allows characterization of these glycopeptides in a glycosylation site-specific fashion. Based on the observed fragmentation patterns, the amount of structural information obtained was dependent on the number of SO_3 groups present and the charge state of the ion. Specifically, it was noted that for all doubly charged ions, when two SO_3 groups were present, the loss of SO_3 was observed as the base peak, and one or two glycan related cleavages were also observed, confirming these ions as sulfated glycopeptides. However, the abundant loss of SO_3 group from the precursor ion resulted in the loss of information about the position of sulfation on the glycopeptide. On the other hand, when one SO_3 group was present, no loss of SO_3 was observed; instead, a loss of CO_2 was observed as the base peak. In this case, no conclusive information was obtained to confirm these ions as sulfated glycopeptides. These results imply that more than one SO_3 group must be present in order to confirm the presence of the SO_3 group for doubly charged sulfated glycopeptides undergoing MS/MS fragmentation in the negative ion mode. While changing the number of

sulfates has a significant effect on the fragmentation of the glycopeptides, changing the peptide present had no observable effect on the MS/MS data.

The amount of structural information obtained was also found to be dependent on the charge state of the ions. It was observed that when monosulfated ions were doubly charged, the loss of SO₃ group was not observed and no useful structural information was obtained. Moreover, for the singly charged state of the same ion, loss of SO₃ group was observed during MS/MS, and substantial structural information was obtained. Based on this comparison, it is evident that the singly charged ions give more structural information than their doubly charged counterparts. This implies that the presence of the singly charged ions is necessary in order to obtain useful structural information for monosulfated glycopeptides. Unfortunately most of the ions observed were doubly charged ions, as shown in Figure 2-1. Therefore if all the glycopeptide compositions assigned contained only one SO₃ group, it would be impossible to obtain any useful information on the doubly charged form of these species. (Observation of singly charged species is limited by the upper mass limit of the instrument). This limits the overall structural information obtained from this approach.

2.3.4 New Approach: The use of ion-pairing with MS/MS

The information obtained from (-) ESI-MS/MS for the sulfated glycopeptides described above can be enhanced by performing MS/MS on

the ion-pair complexes of the glycopeptides. To obtain ion-pairing MS/MS data, in the positive ion mode, the sulfated glycopeptides were combined with Lys-Lys-Lys (3K) in solution, forming an ion-pair complex. This ion-pair complex results from the non-covalent interaction of the SO_3 group present in the glycopeptide with a basic peptide, as shown in Figure 2-3. This interaction stabilizes the SO_3 group thereby promotes dissociation pathways in the positive ion mode that are significantly different than those observed in the negative ion mode, during MS/MS experiments. The basic tripeptide, 3K, was selected as the ion-pairing reagent, since it has proven to be the most effective peptide for binding to sulfated glycans and glycopeptides.²⁵ Table 2-2 represents all the deprotonated doubly charged ions observed in (-) ESI-MS along with the expected m/z values of their corresponding ion-pair complexes that could be observed, when 3K is combined with the eTSH digest. The ions in bold show which ion-complexes were detected after incorporating the ion-pairing reagent.

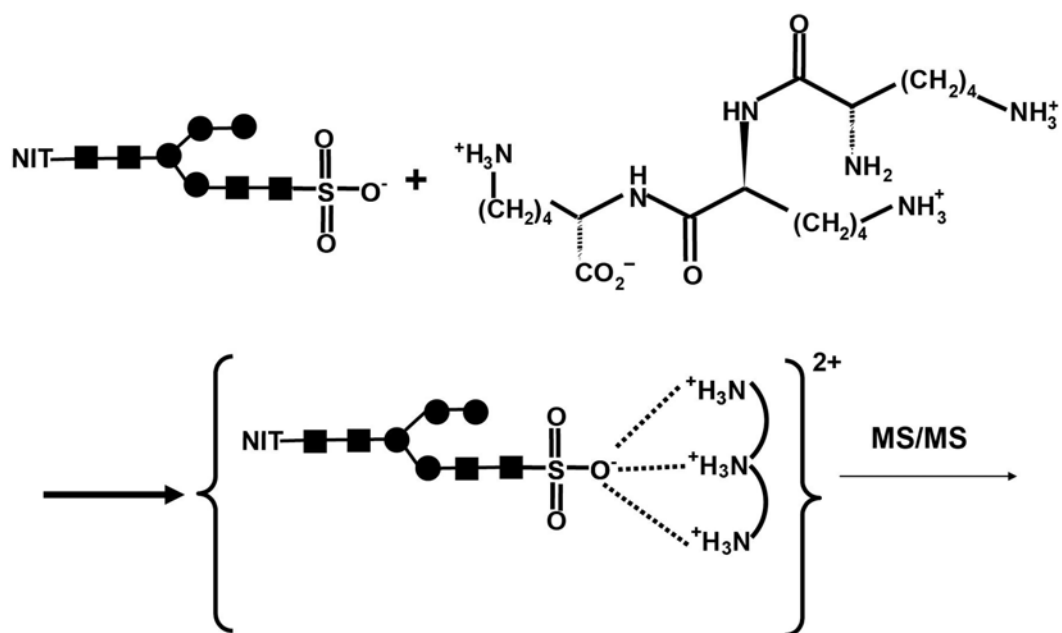


Figure 2-3. Formation of ion-pair complexes for MS/MS analysis.

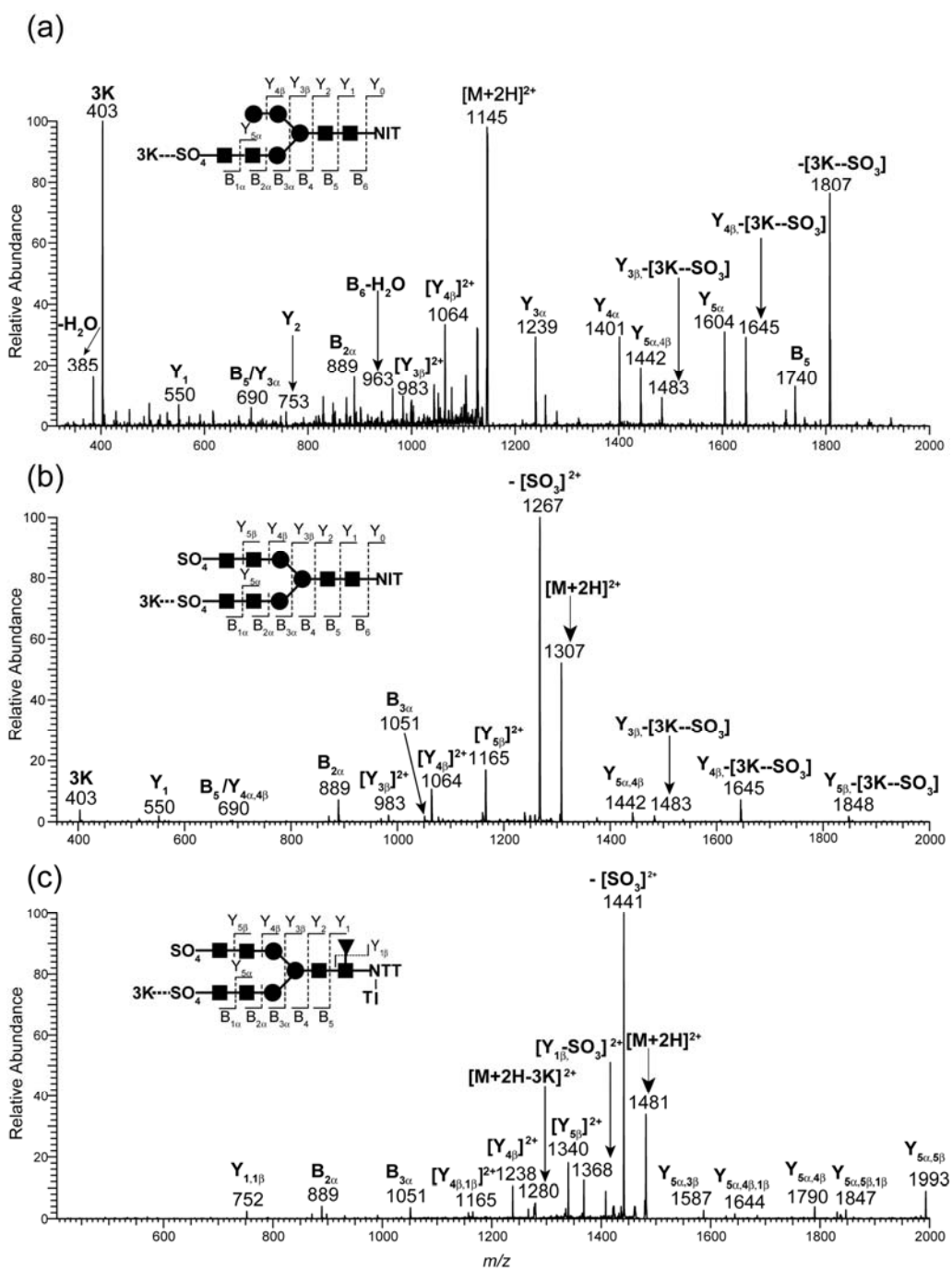


Figure 2-4. Comparison of MS/MS data for ion-pair complexes with Lys-Lys-Lys (3K). (a) MS/MS of a doubly charged monosulfated glycan from N⁵⁶IT glycosylation site; (b) MS/MS of doubly charged disulfated glycan from N⁵⁶IT glycosylation site; (c) MS/MS of a doubly charged disulfated glycan from TIN²³TT glycosylation site.

2.3.5 Comparison of MS/MS data of ion-pair complexes based on

2.3.5.1 Number of SO₃ groups present

As explained earlier, MS/MS data of monosulfated doubly charged ion in Figure 2-2(a) (m/z 942) did not provide useful structural information. However, after the incorporation of the ion-pairing reagent and performing MS/MS in the positive ion mode, more informative product-ions were observed. See Figure 2-4(a). Most of the prominent ions were a result of one or more B- or Y-type glycosidic cleavages. These can be used to verify the composition as well as to provide information on the sequence and branching of the glycan structure. In Figure 2-4(a), $B_{2\alpha}$ (m/z 889) $[Y_{3\beta}]^{2+}$, (m/z 983) and $[Y_{4\beta}]^{2+}$ (m/z 1064) ions not only confirm the presence of the SO₃ group but also indicate its location on the terminal [HexNAc-HexNAc] portion of the glycopeptide. This information was not available before adding an ion-pairing reagent (Figure 2-2(a)).

The pattern of product ions obtained after the addition of the ion-pairing reagent can also be used to define the class of N-linked glycan present. For example, in Figure 2-4(a), the presence of ions, $Y_{3\alpha}$ (m/z 1239) and $Y_{4\alpha}$ (m/z 1401) eliminate the possibility of the N-glycan class being a complex-type glycan, since these ions must be from a high-mannose branch; whereas the $[Y_{3\beta}]^{2+}$ (m/z 983) ion identifies the remaining branch of the glycopeptide that contains the SO₃ group. Therefore, this glycopeptide is

definitely a bi-antennary hybrid structure. This is also supported by the sequential order of Y- and B- ions of the other prominent peaks in the spectrum. Identifying the type of N-linked glycan present is important because different types of these N-linked glycans can have different biological functions.

In addition to obtaining information about the glycan structure, this MS/MS data is also useful to confirm the composition of the peptide. The presence of the Y₁ ion denotes the glycosylation site, because this ion corresponds to [HexNAc+NT]. This ion, when present in MS/MS spectra of ion-pair complexes, can be used to identify the peptide moiety. While the Y₁ ion's presence can be used to confirm the composition of the peptide, when a likely peptide sequence is hypothesized, it would be difficult to identify which ion in the spectrum would correspond to the Y₁ ion, had the peptide composition been completely unknown.

The information obtained from these results collectively allow the full characterization of doubly charged monosulfated glycopeptide by providing structural information on the glycan portion (sequence, branching, and type of N-linked glycan present), confirming the peptide moiety, and confirming the presence/location of the SO₃ group. Similar monosulfated ion-pair complexes fragmented in a similar manner. See supplemental Figure 2-2.

Disulfated glycopeptides were also investigated. These glycopeptides, like their monosulfated counterparts, exclusively form 1:1

complexes with the ion-pairing reagents. The MS/MS data of one of the complexes that contains a disulfated glycopeptide is in Figure 2-4(b). It shows a fragmentation pattern that is similar to the complex in Figure 2-4(a). The main difference is the loss of the SO₃ group, which is the base peak in Figure 2-4(b). This ion is not observed in Figure 2-4(a). This difference is due to the fact that the single SO₃ group present in the mono-sulfated ion is already non-covalently linked to the basic peptide, 3K, which stabilizes it and thus minimizes the loss of SO₃ in the monosulfated ion-pair complexes. This observation is further supported by the appearance of the ions *m/z* 889, 1051, and 983 (doubly charged) corresponding to HexNAc₂+SO₃⁻-3K, HexNAc₂+SO₃⁻-3K, and NIT+HexNAc₄Hex₂+SO₃⁻-3K, respectively (Figure 2-4(b)). The presence of these ions clearly indicates that covalent bonds were broken preferentially, instead of cleaving the non-covalent bond between the SO₃ group and 3K.

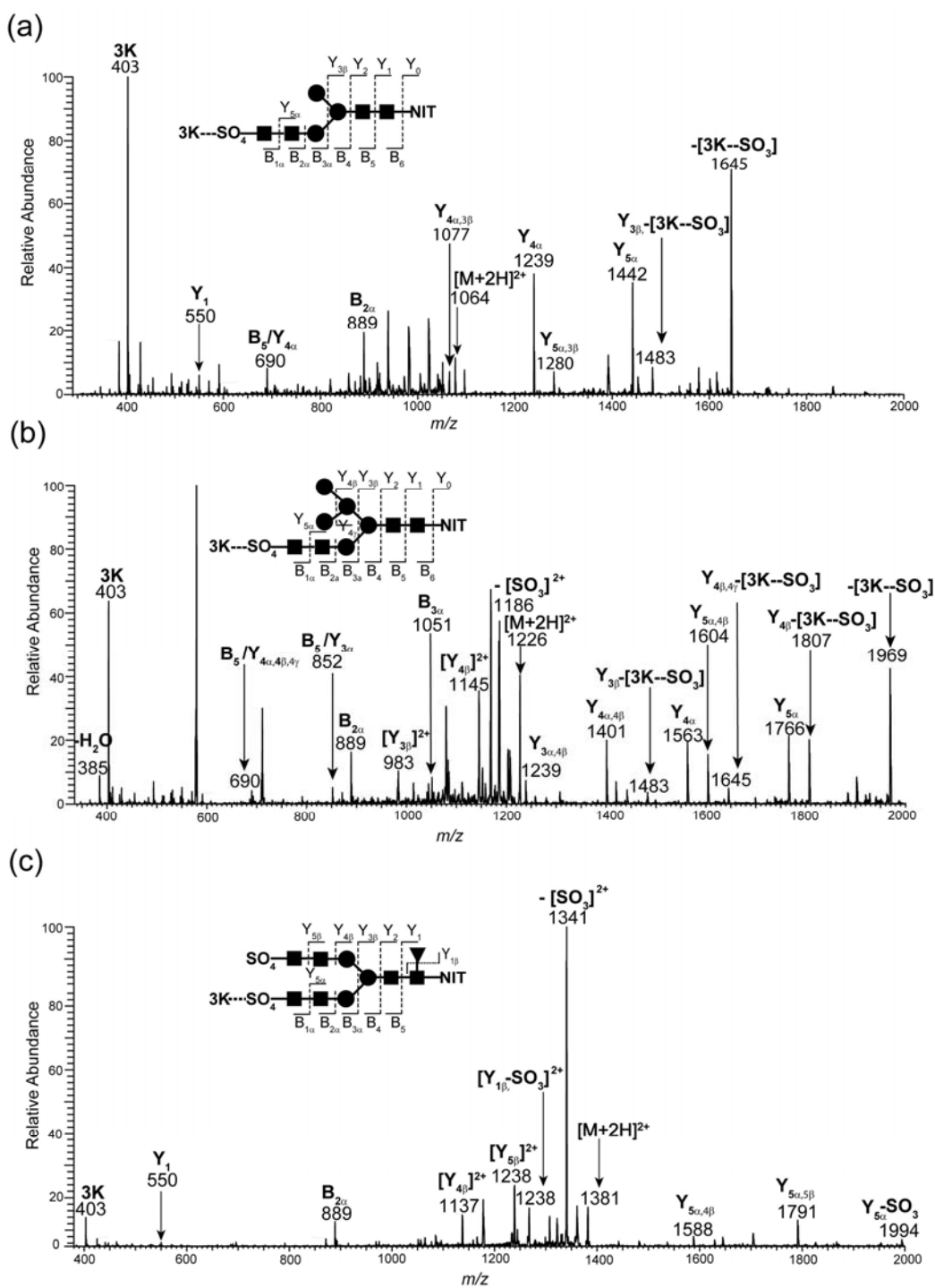
The results obtained in the spectra in Figure 2-4(a) and (b) show that an ion corresponding to loss of SO₃ is the only significant difference in the fragmentation pattern of the glycopeptides with varying numbers of SO₃ groups present. However, when the ion-pairing data is compared to the (-) MS/MS data, a dramatic increase in structural information was observed. The product ions that were observed from ion-pair complexes not only identified the peptide moiety but also confirmed the glycan moiety

composition, the actual position of the SO₃ group, the sequence, branching pattern, and the type of N-linked glycan present.

2.3.5.2 Different glycosylation sites

Since it has been demonstrated that under certain conditions, the peptide moiety in a glycopeptide drives the fragmentation pathways during MS/MS experiments,^{25,26} it is important to determine the effects of the peptide on the fragmentation of these species. Our results show that there is no observable change in fragmentation patterns in (-) ESI-MS/MS data, when the peptide was different (Figure 2-2(b) and (c)). MS/MS data of the same ions after the addition of an ion-pairing reagent in Figure 2-4(b) and (c) again indicates that changes in the peptide composition (from NIT to TINTT) did not affect the mass spectral data. Both spectra are dominated by Y- and B-type ions, providing sequence and branching information of the glycans present. As illustrated before, these ions can also be used to confirm the composition and the type (complex) of N-glycans present. Since the glycopeptide ions, shown in Figure 2-4(b) and (c), have similar glycan moieties, certain ions, such as the B ions (m/z 889 and 1051) appear in both spectra. These ions can be used to infer the location of the SO₃ group (on HexNAc-HexNAc) within the glycopeptide. The Y₁ ion, corresponding to m/z 550 in Figure 2-4(a)) and m/z 752 in Figure 2-4(b), identifies the peptide moieties found on these glycopeptides as NIT and TINTT, respectively.

These results imply that changing the peptide has no significant effect on the fragmentation patterns of the attached sulfated glycans observed in MS/MS experiments. Thus characterization of these glycans in a glycosylation site-specific manner is possible without the peptide moiety complicating the mass spectra. Additional MS/MS data of other ion-pair complexes observed in this study can be found in Supplemental Figure.



Supplemental Figure: These spectra support the fact that the fragmentation trends for the ion-pair complexes described herein of can be generalized to other sulfated glycopeptides.

2.3.6 Application to unknown glycopeptides

In the mass spectral data in Figure 2-1, two prominent ions appear at m/z 1186 and m/z 1228. The compositions of these ions do not appear in Table 2-1, because a logical glycopeptide structure could not be assigned, using peptide sequences from the Edman data in combination with previously-characterized glycans for this hormone. To demonstrate the utility of the method presented herein, and to more fully characterize this glycoprotein sample, both these ions were subjected to (-) ESI-MS/MS experiments and ion-pairing MS/MS experiments.

2.3.6.1 Example 1: Use of MS/MS to differentiate two isobaric structures

In the first example (Figure 2-5), two possible glycan structures are presented. One of the glycans (structure 1) has not been identified previously, and one of them (structure 2) has been previously described. The measured mass that corresponds to the ion is m/z 1186.8856, while the calculated m/z obtained for these species is m/z 1186.8852 for structure 1 and m/z 1186.8859 for structure 2. Both have a mass difference of less than 1 ppm from the measured mass. Since they are isobaric structures, the high-resolution MS data could not discriminate them. Both structures are biologically relevant, so it is important to conduct further analysis in order to confirm the correct structure. The peptide portions (in both examples 1 and 2) of these glycopeptides, though not identified by Edman data, are possibly

present, since both of these peptide sequences represent shorter versions of peptides that have been identified by the Edman data.

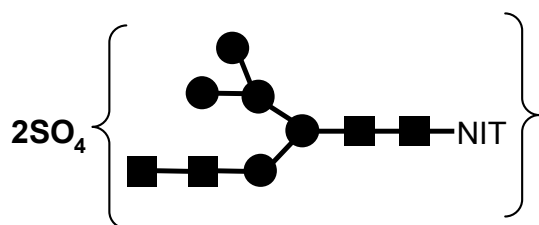
MS/MS in negative ion mode

MS/MS experiments were performed in the negative ion mode to determine which of the two possible structures in Figure 2-5 best corresponds to m/z 1186.8856. The product ions are in Table 2-3A; they were compared against all the possible fragment ions that would be generated by the two structures proposed in Figure 2-5. The loss of SO_3 from the precursor ion to generate the base peak, corresponding to $[\text{M}-2\text{H}-\text{SO}_3]^{2-}$, suggests that this is a disulfated glycopeptide. Since both structure 1 and structure 2 are disulfated, the presence of this ion could not be used to discriminate between the two species. While no logical fragmentation assignments matched m/z 593 and 890 for structure 1, these ions could be assigned to structure 2, as $^{0,2}\text{X}$ and $^{0,2}\text{A}$ ions, respectively, indicating structure 2 as the correct structure. These ions identify the peptide moiety as LENH. This data suggests that structure 2 is the correct structure.

MS/MS of ion-pair complex

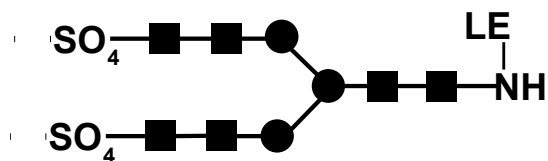
The product ions obtained from MS/MS conducted on the ion-pair complex are summarized in Table 2-3B. Several ions observed in the table were consistent with only one of the two structures. For example the product ion at m/z 715 corresponds to peptide moiety, LENH, attached to GlcNAc and is consistent with the peptide identified in the negative ion mode. Product

ions at m/z 1608, 1649, 1811, and 1891 explicitly indicate that structure 2 is the only possible structure, since these ions exclusively correspond to glycopeptide fragments where the glycan portion is attached to LENH. A few of the ions could not be used to verify the glycopeptide composition, but they provided useful structural information about the location of the SO_3 group within the glycopeptide. For example, product ions like m/z 889 and 1051 verify that the SO_3 group is attached to a terminal [HexNAc-HexNAc] portion of the glycopeptide. These ions could be deduced from either structure 1 or 2, so they were not used to discriminate between the two structures.



Structure 1

Observed m/z : 1186.8856
 Calculated m/z : 1186.8852
 Mass error: 0.3370 ppm.



Structure 2

Observed m/z : 1186.8856
 Calculated m/z : 1186.8859
 Mass error: 0.2528 ppm.

Figure 2-5. Two glycopeptide candidates that could correspond to m/z 1186.

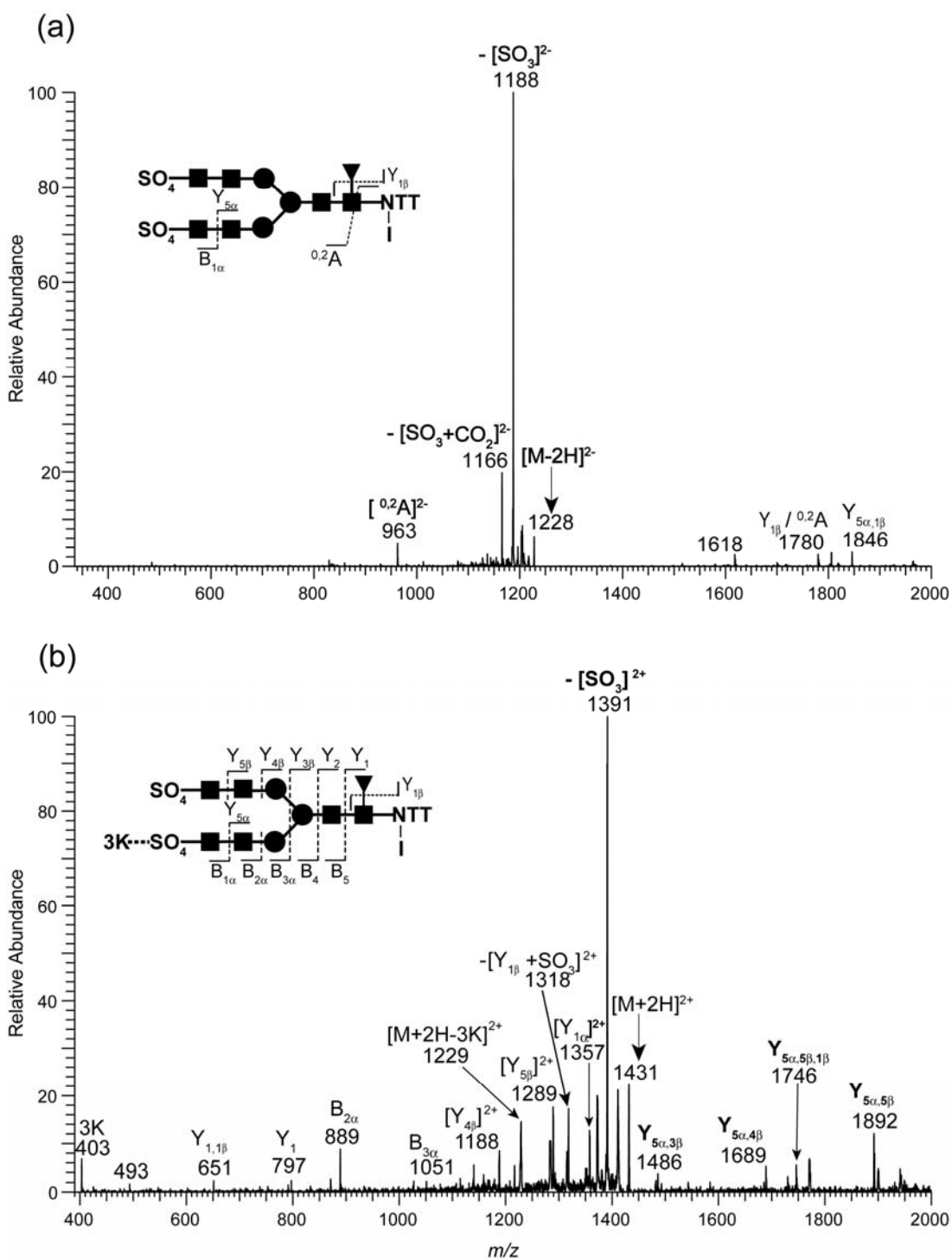


Figure 2-6. MS/MS of an “unknown” glycopeptide (a) in negative ion mode (b) in positive ion mode, after adding an ion-pairing reagent.

Table 2-2: Ion-pair complexes of sulfated glycopeptides that were first identified from (-)ESI-MS data

Glycopeptide	$[M-2H]^{2-}$	M	$[3K+M+2H]^{2+}$	$[3K+M+3H]^{3+}$
1	861.3	1724.6	1064.5	710.0
2	942.3	1886.7	1145.5	764.0
3	1023.4	2048.7	1226.5	818.0
4	1064.4	2130.8	1267.5	845.4
5	1104.4	2210.7	1307.5	872.0
6	1177.4	2356.8	1380.6	920.7
Unknown 1	1186.9	2375.8	1390.0	927.0
Unknown 2	1227.9	2457.8	1431.1	954.4
7	1278.4	2558.9	1481.6	988.1

$[M-2H]^{2-}$ represents the doubly charged ions' m/z observed in (-)ESI-MS. M is the molecular mass of the glycopeptide before adding the ion-pairing reagent; 3K = ion-pairing reagent, which has a mass of 402.3 Da. $[3K+M+2H]^{2+}$ and $[3K+M+3H]^{3+}$ correspond to the m/z for the doubly and triply charged ion-pair complexes that could be observed in positive ion mode. Peaks in bold correspond to complexes that were detected.

Table 2-3: Identification of the correct structure for m/z 1186 by MS/MS
A) Product ions from (-)MS/MS

Product ion Observed (m/z)	Assignment	Product ions consistent with Structure 1	Product ions consistent with Structure 2
1146	$[M-2H-SO_3]^{2-}$	Yes	Yes
593	$[^{0,2}X]^{-1}$ from Structure 2	No	Yes
890	$[^{0,2}A]^{2-}$ from Structure 2	No	Yes

B) Product ions from MS/MS of the ion-pair complex

Product ion Observed (m/z)	Assignment	Product ions consistent with Structure 1	Product ions consistent with Structure 2
889	HexNAC ₂ + SO ₃ ---3K	Yes	Yes
1051	HexNAC ₂ Hex + SO ₃ ---3K	Yes	Yes
715	LENH+HexNAC	No	Yes
1608	LENH+HexNAC ₃ Hex ₃	No	Yes
1649	LENH+HexNAC ₄ Hex ₂	No	Yes
1811	LENH+HexNAC ₃ Hex ₃	No	Yes
1891	LENH+HexNAC ₃ Hex ₃	No	Yes

By comparing the data in Tables 2-3A and B, it is clear that both MS/MS methods are useful in obtaining the peptide composition for this particular glycopeptide. These data also clearly show that the ion-pairing approach can not only be used to verify the glycan composition, but it can also be used to infer the location of the sulfate within the glycopeptide.

2.3.6.2 Example 2: Use of MS/MS analysis for structural information

MS/MS in negative ion mode

Figure 2-6 (a) represents (-) MS/MS of another unknown glycopeptide, m/z 1228. Based on the fragmentation patterns observed in the negative ion mode, the presence of $^{0,2}A$ ion at m/z 963 identifies the peptide moiety as INTT, and the apparent loss of SO_3 from the precursor ion to generate the base peak suggests that this is a disulfated glycopeptide. Using this information, a glycopeptide structure can be proposed that is consistent with the high-resolution data, the peptide assignment (INTT), and previously characterized glycan moieties. This structure is shown on the spectrum in Figure 2-6(a). The presence of $Y_{5\alpha,1\beta}$ at m/z 1846 identifies the loss of fucose, along with the loss of SO_3 and terminal HexNAc, whereas the presence of $Y_{1\beta} / ^{0,2}A$ ion at m/z 1780 further confirms the peptide moiety and the likely position of the fucose at the core HexNAc. Unfortunately, this information is inadequate to completely verify the structure of the glycan.

(+)MS/MS of ion pair complex

Figure 2-6 (b) represent MS/MS of the ion-pair complex of the same glycopeptide shown in Figure 2-6 (a). The presence of Y_1 and $Y_{1,1\beta}$ ions at m/z 797 and 651 identify the peptide as INTT. These ions also confirm the presence of a fucose, and where it is located (on the core HexNAc). Presence of $B_{3\alpha}$ and $B_{2\alpha}$ ions at m/z 1051 and 889 identify the location of SO_3 on the terminal [HexNAc-HexNAc]. These ions also identify the presence of a complex branch (HexNAc-HexNAc- SO_3+3K) in this glycopeptide. The loss of SO_3 from the precursor ion, as the base peak, confirms this to be a disulfated glycopeptide. Assuming triamannosyl core is intact, the abundant glycosidic cleavage ions can be used to infer the structure. Since fucose is the only monosaccharide that is lost without a concerted loss of sulfate, fucose must be the only terminating monosaccharide. This implies that the two sulfate groups are each capping a HexNAc-HexNAc disaccharide branch, as depicted in the figure. All the observed glycosidic cleavage ions support this assignment. Assuming this glycopeptide was a complete unknown, this information can explicitly identify and characterize the unknown sulfated glycopeptide: Not only was the sequence, branching pattern, and the type of glycan identified, but also the peptide and the number of SO_3 groups present were determined.

From the two MS/MS approaches, it is evident that the use of MS/MS in the negative ion mode is the better approach to identify the peptide moiety of unknown glycopeptides, when Edman data is unavailable or when sample

quantity is limited. This is because, unlike in the positive ion mode, only one or two peaks of lower mass are observed. However, MS/MS data of ion-pair complexes are very useful in identifying the location of SO₃, the composition, the sequence, branching and the type of N-linked glycans present.

2.4 Conclusion

The results from this study clearly demonstrate the efficacy of using ion-pairing MS/MS to fully characterize sulfated glycopeptides in a glycosylation site-specific fashion, an approach that is complementary and in most cases superior to MS/MS analysis in negative ion mode. While the negative mode MS/MS data was useful at identifying the peptide moiety, it provided minimal glycan structural information. This information was dependent on the number of SO₃ groups and the charge state of the ion. In contrast, the ion-pairing method provided a wealth of structural information about the glycan portion in addition to being useful for identifying the peptide moiety. The information obtained from MS/MS of the ion-pair complexes can be used to determine the branching, sequence, and type of N-glycan present in a sulfated glycopeptide. This is because MS/MS of the corresponding ion-pair complexes provided several glycosidic cleavage ions. This information was readily available from all ion-pair complexes observed, regardless of the number of SO₃ groups present or the charge state of the ion.

2.5 References

1. Hooper, L. V.; Manzella, S. M.; Baenziger, J. U. *FASEB J.* **1996**, 10, 1137-1146.
2. Hortin, G.; Green, E. D.; Baenziger, J. U.; Strauss, A. W. *Biochem J.* **1986**, 235, 407-414.
3. Van Rooijen, J. J. M.; Kamerling, J. P.; Vliegenthart, J. F.G. *Eur J Biochem* **1998**, 256, 471-487.
4. Plaas, A. H.; West, L. A.; Wong-Palms, S.; Nelson, F. R. *J Biol Chem.* **1998**, 273, 12642-12649.
5. Bayliss, M. T.; Osborne, D.; Woodhouse, S.; Davidson, C. *J Biol Chem.* **1999**, 274, 15892-15900.
6. Tondorov, P. T.; Deacon, M.; Tisdale, M.J. *J Bio Chem.* **1997**, 272, 12279-12288.
7. Thomsson, K. A.; Karlsson N, G.; Hansson, G. C. *J Chromatogr A.* **1999**, 854, 131-139.
8. Sangadala, S.; Bhat, U. R.; Mendicino, J. *Mol Cell Biochem.* **1993**, 126, 37-47.
9. Bowman, K. G.; Cook, B. N.; de Graffenried, C. L.; Bertozzi, C. R. *Biochemistry* **2001**, 40, 5382-5391.
10. Green, E. D.; Baenziger, J. U.; Boime, I. *J Biol Chem.* **1985**, 260, 15631-15638.
11. Green, E. D.; Baenziger, J. U. *J. Bio Chem.* **1988**, 263, 36-44.
12. Parsons, T. F.; Pierce, J.G. *Proc Natl Acad Sci U S A.* **1980**, 77, 7089-7093.
13. Hortin, G.; Natowicz, M.; Pierce, J.; Baenziger, J.; Parsons, T.; Boime, I. *Proc Natl Acad Sci U S A.* **1981**, 78, 7468-7472.

14. Fiete, D.; Srivastava, V.; Hindsgaul, O.; Baenziger, J. U. *Cell*. **1991**, 67, 1103 - 1110.
15. Delcommenne, M.; Kannagi, R.; Johnson, P. *Glycobiology* **2002**, 12, 613-622.
16. Gesundheit, N.; Gyves, P. W.; DeCherney, G. S.; Stannard, B. S.; Winston, R. L.; Weintraub, B.D. *Endocrinology*. **1989**, 124, 2967-2977.
17. Sundblad, G.; Kajiji, S.; Quaranta, V.; Freeze, H. H.; Varki, A. *J Biol. Chem.* **1988**, 263, 8897-8903.
18. Gesundheit, N.; Magner, J. A.; Chen, T.; Weintraub, B. D. *Endocrinology* **1986**, 119, 455-463.
19. Zaia, J. *Mass Spectrom Rev.* **2004**, 23, 161-227.
20. An, H. J.; Peavy, T. R.; Hedrick, J. L.; Lebrilla, C. B. *Anal Chem.* **2003**, 75, 5628-5637.
21. Taguchi, T.; Iwasaki, M.; Muto, Y.; Kitajima, K.; Inoue, S.; Khoo, K. H.; Morris, H. R.; Dell, A.; Inoue, Y. *Eur J Biochem.* **1996**, 238, 357-67.
22. Wheeler, S. F.; Harvey, D. J. *Anal Biochem.* **2001**, 296, 92-100.
23. Wuhler, M.; Koeleman, C. A. M.; Hokke, C. H.; Deelder, A. M. *Anal Chem.* **2005**, 77, 886-894.
24. Jiang, H.; Irungu, J.; Desaire, H. *J Am Soc Mass Spectrom.* **2005**, 16, 340-348.
25. Jiang, H.; Desaire, H.; Butnev, V. Y.; Bousfield, G. R. *J Am Soc Mass Spectrom.* **2004**, 15, 750-758.
26. Conboy, J. J.; Henion, J. D. *J Am Soc Mass Spectrom.* **1992**, 3, 804-814.
27. Juhasz, P.; Biemann, K. *Proc. Natl. Acad. Sci. USA* **1994**, 91, 4333-37.
28. Siegel, M. M.; Tabei, K.; Kagan, M. Z.; Vlahov, I. R.; Hileman R.E.; Linhardt, R. J. *J. Mass Spectrom.* **1997**, 32, 760-772.

29. Gunay, N. S.; Tadano-Aritomi, K.; Toida, T.; Ishizuka, I.; Linhardt, R. J. *Anal. Chem.* **2003**, 75, 3226-3231.
30. Kuberan, R.; Lech, M.; Zhang, L.; Wu, Z. L.; Beeler, D. L.; Rosenberg, R. D. *J. Am. Chem. Soc.* **2002**, 124, 8707-8718.
31. Zhang, Y.; Go, E. P.; Jiang, H.; Desaire, H. *J. Am. Soc. Mass Spectrom.* **2005**, 16, 1827-1839.
32. Bousfield, G.R; Ward, D.N. *J Biol. Chem.* **1984**, 259, 1911-1921.
33. Wurzel, C.; Wittmann-Liebold, B. *EXE.* **2000**, 88, 145-157.

CHAPTER 3

Simplification of mass spectral analysis of acidic glycopeptides using *GlycoPep ID*.

Reprinted by permission from *Anal Chem.* **2007**, 79, 3065-3074. Copyright 2007 American Chemical Society

3.1 Introduction

Glycosylation is a ubiquitous post-translational modification that occurs in most secreted and membrane-bound proteins.¹⁻⁴ This process is known to influence various structural and functional properties of the glycosylated proteins. For example, glycosylation affects protein folding,^{1,2} solubility,^{3,5} antigenicity,^{1,6-8} biological activity,^{2,3} half-life in circulation^{1,2,9} and is also important in protein-carbohydrate interactions.^{1,5} Changes in glycosylation have been implicated in several diseases such as cancers¹⁰ and congenital disorders.¹¹ Some of these modifications involve sulfation and/or sialylation of the attached glycan moiety, and these groups transform the modified protein into a unique structure with different biological properties. Because these modifications are of eminent biological significance, their analysis is the focus of this study.

A thorough analysis of glycosylation on proteins involves characterizing the composition of the glycans present at each glycosylation site. When only one glycosylation site is present on a protein, the carbohydrates can be cleaved from the protein - either chemically or

enzymatically - and characterized separately.¹² However, this approach has limited utility when more than one glycosylation site is present on a protein, because information about which carbohydrates originated from which glycosylation sites is lost prior to analysis. To retain this information, a different analysis strategy must be employed. Glycosylated proteins can be analyzed by digesting the protein and characterizing glycopeptides by mass spectrometry.^{2-8,13-40} This approach is advantageous because once the structure is characterized, the identified peptide portion of the glycopeptide can be used to determine *where* the carbohydrate is attached on the protein. As described below, this strategy needs some development before it can be implemented in a routine fashion.

When proteolytic digestion is used for the analysis of glycopeptides, most studies usually employ trypsin as the protease, since its cleavage sites are at specific amino acid residues, and the sites are well known. Thus, the peptide sequences that are components of the glycopeptides are readily predictable, provided a protein sequence is available. Since all the possible peptide sequences are known *a priori*, assigning peptide compositions to the glycopeptides is fairly straightforward, and it can be accomplished using a variety of approaches. However, proteolysis of the glycoprotein using trypsin has several limitations. Detecting the glycosylated peptides, which have lower ionization efficiency than nonglycosylated peptides, is one major obstacle. Additionally, missed tryptic cleavages, which are known to occur,

significantly complicate data analysis.^{5-8,12} Finally, some glycosylated proteins are not highly susceptible to tryptic digest, and others contain glycosylation sites that are near each other; thus they yield glycopeptides with more than one glycosylation site on the same glycopeptide.^{7,8,20,37} For example, Cutalo et al. could not characterize five glycosylation sites in gp120, an HIV envelope glycoprotein, due to incomplete trypsin digestion and the presence of multiple glycosylation sites within one peptide.⁷ This prevented a complete glycosylation site-specific analysis of the attached glycans.

To overcome the limitations described above, several researchers have begun using digestion enzymes like proteinase K or pronase; which proteolyze proteins in a non-specific fashion.^{17-20,25} Using these non-specific enzymes result in a mixture of glycopeptides that contain significantly shorter peptide sequences, and these glycopeptides are less challenging to detect by MS analysis for a couple of reasons. First, sections of the protein that are not glycosylated become almost completely digested, so non-glycosylated peptides do not appear in the same mass range as the glycopeptides.²⁵ Therefore, these two species can be readily distinguished from each other. In addition, the hydrophobicity of the small nonglycosylated peptides that are present after the digest are quite different from that of glycopeptides, so they can be rapidly separated from glycopeptides using a variety of fractionation strategies. While digestion with non-specific enzymes provides a clear

advantage over trypsin in terms of data acquisition, it can be almost impossible to identify the peptide portion of these glycopeptides. This is because the amino acid residues present in the peptide portions of glycopeptides generated from a non-specific enzyme like proteinase K can vary greatly; they can contain two to five amino acid residues including the N-glycosylation site.³⁸ As a result, efficient and reliable methods are required to explicitly identify the peptide portion for these glycopeptides.

So far, only a few methods are currently available for helping to identify the peptide portion when glycoproteins are proteolyzed in a non-specific manner. For example, when pronase or proteinase K is used to produce glycopeptides, the peptide portion can be identified by either comparing the mass of the glycosylated peptide with the nonglycosylated peptide (after being subjected to endoglycosidase digestion)^{20,25} or by using Edman chemistry.^{17-19,39} The former method is less reliable than Edman chemistry, since it identifies the peptide based on the mass difference, and hence it is not a feasible solution when analyzing complex mixtures. Edman chemistry's advantage is that the exact amino-acid sequences can be determined, but this approach requires the glycopeptides to be isolated from the nonglycosylated components before analysis, and Edman data from complex glycopeptide mixtures is also very difficult to interpret. Finally, this technique supplies aggregate data about all the peptide sequences that are present, but it does not link the peptide sequence specifically to each of the

glycopeptides detected during MS analysis. Both Edman Chemistry and endoglycosidase digestion also suffer the limitation that additional chemical reactions need to be performed to obtain the peptide information, which means more sample and more experiment time is required. In positive ion mode, ECD is also helpful, but this requires that the glycopeptides are at least in the +2 charge state, so these methods are not feasible for acidic glycopeptide analysis.⁴¹ Consequently, an alternative approach for peptide identification during glycopeptide analysis is highly desirable.

Herein, we propose a novel approach that is facile and highly effective in identifying the peptide moiety of glycopeptides generated using a non-specific enzyme. The approach utilizes our newly developed web-based program, GlycoPep ID. This program determines the peptide portion of glycopeptides by calculating the theoretical m/z of $^{0,2}X$ ion that can be generated from a glycoprotein of interest. GlycoPep ID will generate a table of the predicted peptide sequences with their corresponding m/z values, and a list of predicted m/z 's of the $^{0,2}X$ ions that would be generated for each peptide. The $^{0,2}X$ ion is listed in the table because this ion has been shown to be a characteristic for acidic glycopeptides undergoing CID.¹⁹ Once the predicted $^{0,2}X$ ions are obtained, product ions present in CID experiments are searched against this table to locate a match. Matched ions identify the peptide portion of the glycopeptide. The versatility of this method is demonstrated in the identification of the correct peptide moieties of

previously characterized sulfated glycopeptides from a proteinase K digest of equine thyroid stimulating hormone (eTSH). This glycoprotein consists of three glycosylation sites that are heavily glycosylated with a varying degree of sulfation at each site.¹⁹ The peptide identification method is also extended to identifying the peptide moieties of glycopeptides from a more complicated glycoprotein, follicle stimulating hormone (FSH), that consists of four glycosylation sites, each containing various degrees of sialylation, sulfation, or both sialylation and sulfation. All the peptide moieties identified in this study were previously verified with Edman data in conjunction with high-resolution FTICR-MS analysis. The present strategy has several advantages over the previously used approaches for identifying the peptide portions for these glycopeptides. It has lower sample consumption and analysis time, but more importantly composition information about every glycopeptide present in a complex mixture can be obtained in a single series of CID experiments.

3.2 Experimental/Methods

3.2.1 Enzymatic digestion of eTSH and eFSH with Proteinase K

Proteolytic digestion of equine thyroid stimulating hormone (eTSH) and follicle stimulating hormone (eFSH) glycoproteins to generate glycopeptides was performed by Dr. George Bousfield of Wichita State University as described in the protocol by Bousfield et. al.⁴² Briefly, each glycoprotein was reduced, alkylated and desalted before digesting with proteinase K.⁴² The dried digests from eTSH and eFSH were then subjected to Superdex peptide gel filtration chromatography.^{19, 39} The glycopeptide fraction was collected, dried, and analyzed as described below.

3.2.2 Glycopeptides preparation for mass spectrometry analysis

The dried glycopeptide sample was dissolved in water and diluted with MeOH:H₂O (4:1) containing 0.3% acetic acid, to constitute a final concentration of 0.03 µg/µL. This solution was introduced into the mass spectrometer by direct infusion using a syringe pump at a flow rate of 5 µL/min.

3.2.3 CID experiments

CID experiments were performed using electrospray ionization (ESI) in the negative ion mode on a linear ion trap-Fourier Transform Ion Cyclotron Resonance mass spectrometer, LTQ-FTICR MS, (Thermo Finnigan, San Jose, CA) to identify the peptide moiety for each glycopeptide analyzed. All

CID data were acquired on LTQ. The precursor ions were selected with an isolation width of 3 Da and activated for 30 ms. The activation q_z was maintained at 0.25. Activation amplitudes were in the range of 15-29%, as defined by the instrument software. Data was acquired and processed using Xcalibur 1.4 SR1 software (Thermo Finnigan San Jose, CA). The glycopeptide compositions were validated based on previous characterizations of these samples.^{19,39}

3.2.4 Data Analysis

Data from CID experiments were analyzed using our newly developed in-house web-based program, GlycoPep ID. GlycoPep ID provides a web interface (Figure 3-1) that enable the user to specify the experimental parameters and the peak list from the CID data. The input is processed using the user specified settings that are described in the next section. GlycoPep ID generates a table of the predicted peptide sequences, the theoretical peptides' m/z values, and a list of predicted m/z 's of the characteristic signature product ion, $^{0,2}X$, resulting from the fragmentation of glycopeptides. Data analysis was performed by comparing the peak lists obtained from CID experiments with list of theoretical m/z 's of the $^{0,2}X$ ion. When a single match is found, the $^{0,2}X$ ion is used to determine the peptide portion of the glycopeptide. In the case where more than one match is found, the existence of the $^{0,2}A$ ion, the complimentary ion to $^{0,2}X$, is examined in the CID data. Both $^{0,2}X$ and $^{0,2}A$ ions are used to determine the

peptide portion of the glycopeptide. GlycoPep ID calculates the m/z value for $^{0,2}A$ ions by subtracting the mass of a theoretical “ $^{0,2}X$ neutral loss” directly from the precursor ion. As a result, when $^{0,2}A$ ions are to be calculated, the user must input the singly charged form of the precursor ion.

3.3 Results and Discussion

3.3.1 GlycoPep ID Overview

GlycoPep ID is a freely accessible web-based program that we developed specifically to identify the peptide portion of glycopeptides. The glycopeptides may be generated from proteolytic cleavage with either a specific or non-specific enzyme. This manuscript exclusively describes GlycoPep ID's use in identifying the peptide portion of glycopeptides using proteinase K. To identify the peptide portion, the user inputs the protein sequence and the experimental parameters used, which include the enzyme (proteinase K or trypsin), cysteine modification, charge state, and mass tolerance in ppm or Da into the program (Figure 3-1). The user also inputs the extracted peak list from the CID data. An abundance threshold of 10% is set for all ions in the selected mass range to be included in the peak list. The user has the option to calculate the $^{0,2}A$ ion, which is complementary to the $^{0,2}X$ ion.

Once all the necessary parameters are input and submitted, GlycoPep ID first determines how many N-linked glycosylation sites are present in the

protein of interest. Once the glycosylation sites are determined, it generates a table consisting of all possible peptide sequences from these glycosylation sites that range between two to six amino acids in length that includes the glycosylation site. In addition to the peptides that could be present, GlycoPep ID lists theoretical m/z 's for each of the peptides, the corresponding m/z 's of $^{0,2}X$ ion, and plausible matches from CID data input by the user. When a single match is obtained, the peptide portion is identified. When more than one match is obtained, the CID data is searched for the $^{0,2}A$ ion. In this case, the correct match from the output is one that has both $^{0,2}X$ and $^{0,2}A$ ion in the mass spectrum. It should be noted that all masses reported by the program are monoisotopic.

GlycoPep ID was implemented using an open source server side scripting language PHP running under an Apache web server on a Linux system. The program can be accessed from the website, <http://hexose.chem.ku.edu/sugar.php>, under Tools. Future updates of GlycoPep ID will include variable peptide modifications, charge carriers other than H^+ , and the ability to choose other proteases for theoretical digestion.

The screenshot shows the GlycoPep ID web application. At the top, the title "GlycoPep ID" is displayed in a stylized font. On the left side, there is a vertical navigation menu with the following items: "ABOUT GlycoPep ID", "HELP", "TOOLS", "CONTACT INFO", and "HOME". The main content area is titled "Protein Sequence" and contains a large, empty text input field. Below this field, there are several configuration options: "Enzyme:" with a dropdown menu set to "Proteinase K", "Charge:" with a dropdown menu set to "+1", and "Mass Tolerance ±" followed by an empty text input field and a dropdown menu set to "ppm". Below these, there is a "Cysteine Modification" dropdown menu set to "none". The next line asks "Calculate the ^{0,2}A ion?" with radio buttons for "Yes" and "No", where "No" is selected. Below that, it says "If Yes, Specify the singly charged parent ion" followed by an empty text input field. At the bottom of the main area, there is a section titled "PeakList" with another large, empty text input field. At the very bottom of the page, there are two buttons: "Submit" and "Clear".

Figure 3-1: A screenshot of GlycoPep ID.

**AMINO ACID SEQUENCE OF EQUINE THYROID STIMULATING HORMONE
(eTSH)**

Alpha Subunit

FPDGEF TTQDCPECKL RENKYFFKLG VPIYQCKGCC FSRAYPTPAR SRKT
MLVPKN⁵⁶ITSESTCCVA KAFIRVTVMG NIKLEN⁸²HTQC YCSTCYHHKI

Beta Subunit

FCIPTEYMMH VERKECAYCL TIN²³TTICAGY CMTRDINGK LFLPKYALSQD
VCTYRDFMYK TVEIPGCPDH VTPYFSYPVA VSCKCGKCNT DYSDCIHEAI
KANYCTKPQK SYVVEFSI

AMINO ACID SEQUENCE OF EQUINE FOLLICLE STIMULATING HORMONE

(eFSH)

Alpha subunit

FPDGEFTTQDCPECKLRENKYFFKLGVPIYQCKGCCFSRAYPTPARSRKTMLVPK
N⁵⁶ITSESTCCVAKAFIRVTVMGNIKLEN⁸²HTQC YCSTCYHHKI

Beta subunit

NSCELN⁷ITIAVEKEECGFCISIN²⁴TTWCAGYCYTRDLVYKDPARPNIQKTCTFKEL
VY
ETVK VPGCAHHADS LYTYPVATAC HCGKCNDSST DCTVRGLGPS
YCSFGDMKE

Figure 3-2(a): Amino acid sequence of eTSH obtained from Swiss-Prot. Its three glycosylation sites, α Asn⁵⁶, α Asn⁸², and β Asn²³ are highlighted.(b) : Amino acid sequence of eFSH obtained from Swiss-Prot. Its four glycosylation sites, α Asn⁵⁶, α Asn⁸², β Asn⁷, and β Asn²³ are highlighted.

3.3.2 Method validation

The glycoproteins used in this analysis are heterodimeric pituitary hormones, thyroid stimulating hormone (TSH) and follicle stimulating hormone (FSH) containing three and four glycosylation sites, respectively. Each of the glycosylation sites is known to have a wide diversity of N-linked glycans. These glycoproteins have been extensively studied and therefore they are ideal model systems for developing a method for glycopeptide analysis. The protein sequences are shown in Figure 3-2.

In the first set of experiments, the glycoprotein, eTSH, was employed to demonstrate the viability of using GlycoPep ID in identifying the peptide portion of sulfated glycopeptides. Preliminary data from eTSH indicated that (-) ESI-MS/MS of sulfated glycopeptides always produced both $^{0,2}X$ and $^{0,2}A$ ions, which are complementary ions resulting from a cross-ring cleavage of the carbohydrate attached to the peptide as shown in Figure 3-3.¹⁹ The goal of the work presented herein is to use the information that this cleavage occurs readily as the basis of our approach for identifying the peptide portion of any negatively charged glycopeptide, using GlycoPep ID.

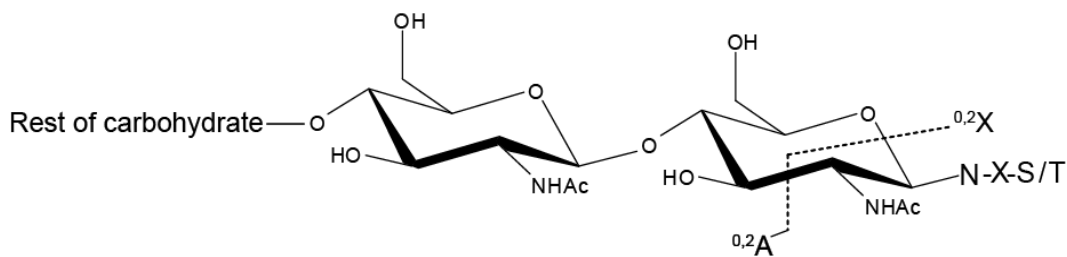


Figure 3-3: The $^{0,2}X$ and $^{0,2}A$ cross-ring cleavage, which occurs during CID experiments of negatively charged glycopeptides. The ions produced by this cleavage are used for identifying the peptide portion of glycopeptides.

(-)ESI-MS/MS data obtained from all the previously identified sulfated glycopeptides of eTSH was analyzed by comparing the product ion masses to the list of predicted masses of the $^{0,2}X$ ion from GlycoPep ID. Figure 3-4a and b shows (-) ESI-MS/MS data of two glycopeptides that originate from two different glycosylation sites in eTSH. In this data, several product ions are present that could potentially correspond to the expected $^{0,2}X$ ion that would identify the peptide moiety. In order to identify the $^{0,2}X$ ion, the data was analyzed using GlycoPep ID. The following parameters were used for GlycoPep ID: the peak list from the (-) ESI-MS/MS data in Figure 3-4a, the eTSH protein sequence, proteinase K for enzyme, charge state of -1, mass tolerance of ± 0.1 Da, carbamidomethyl for cysteine modification. GlycoPep ID outputs the prediction table for eTSH and a plausible $^{0,2}X$ candidate from the peak list (Figure 3-5A). In this analysis, only m/z 630 is identified as a matching ion, as shown in Figure 3-5A. This implies that TIN²³TT is the correct peptide present in this glycopeptide.

A similar interpretation can be made for the mass spectral data in Figure 3-4b. When the peak list of the product ions in Figure 3-4b is input with the following parameters: eTSH protein sequence, proteinase K for enzyme, charge state of -1, mass tolerance of ± 0.1 Da, carbaimidomethyl for cysteine modification, GlycoPep ID outputs more than one plausible $^{0,2}X$ candidate. The matches corresponded to the peptides NIT/TIN/INT (m/z 428.2), LENHT (m/z 694.3), and NITSES (m/z 731.3) respectively within a mass error of ± 0.1 Da. Because more than one match was generated, the data was re-analyzed to calculate the $^{0,2}A$ ion. In this analysis, the singly charged precursor ion at m/z 2355.8 was input, per the requirements of the algorithm. The corresponding prediction table and the list of plausible $^{0,2}X$ and $^{0,2}A$ candidates are shown in Figure 3-5B. Using the masses of the singly charged $^{0,2}A$ ions output from GlycoPep ID, the mass spectrum was inspected to identify either singly charged or doubly charged forms of these ions. Of the three potential peptide sequence ions identified by GlycoPep ID, only m/z 428 has its complementary ion (m/z 963) present in the spectrum. Therefore, the correct peptide match corresponds to either $N^{56}IT$ or $IN^{23}T$. While the MS data cannot discriminate among sequences of isomeric peptides, previous analysis on this sample indicated that all the glycopeptides containing only N, I, and T amino acid residues originated from the alpha subunit (α -Asn⁵⁶). As a result, the correct peptide moiety for this disulfated glycopeptide is identified as $N^{56}IT$.

The peptide moieties for all the other sulfated glycopeptides in eTSH are identifiable in a similar manner. MS/MS data was acquired on each previously characterized eTSH glycopeptide,¹⁹ and a list of all the observed product ions for each glycopeptide was generated. Analysis of this data using GlycoPep ID identified one unique peptide moiety for each glycopeptide analyzed. The results of this analysis for all glycopeptides in eTSH are summarized in Table 3-1. This table contains a total of nine sulfated glycopeptides that are either mono- or di-sulfated. From these sulfated glycopeptides, four different peptide moieties (N⁵⁶IT, TIN²³TT, IN²³TT, LEN⁸²H) were identified; these peptides corresponded to all the three glycosylation sites in eTSH. Although in some cases the peptide sequence could not be identified exclusively, for example, *m/z* 529 in Table 3-1 matches three peptide masses in the prediction table that GlycoPep ID generates, TIN²³T, IN²³TT, and N²³TTI. All these possible peptides map to the same glycosylation site. Since the goal is to determine which carbohydrates are attached to which glycosylation site, this small ambiguity in the peptide sequence is irrelevant in this case. In every case, the peptide identified using this strategy matched the validation data, where a combination of Edman sequencing and FTICR-MS was required to obtain the same information.¹⁹

Based on these results, it is quite evident that (-)ESI-MS/MS data can be used in conjunction with GlycoPep ID to identify the peptide moiety

present in sulfated glycopeptides that are generated by proteinase K digestion, and this approach has significant advantages over the current methods of obtaining this information. This method is more sensitive than Edman sequencing, and it is a more selective technique, since it provides information on each glycopeptide specifically, instead of simply supplying aggregate information on the entire glycopeptide mixture.

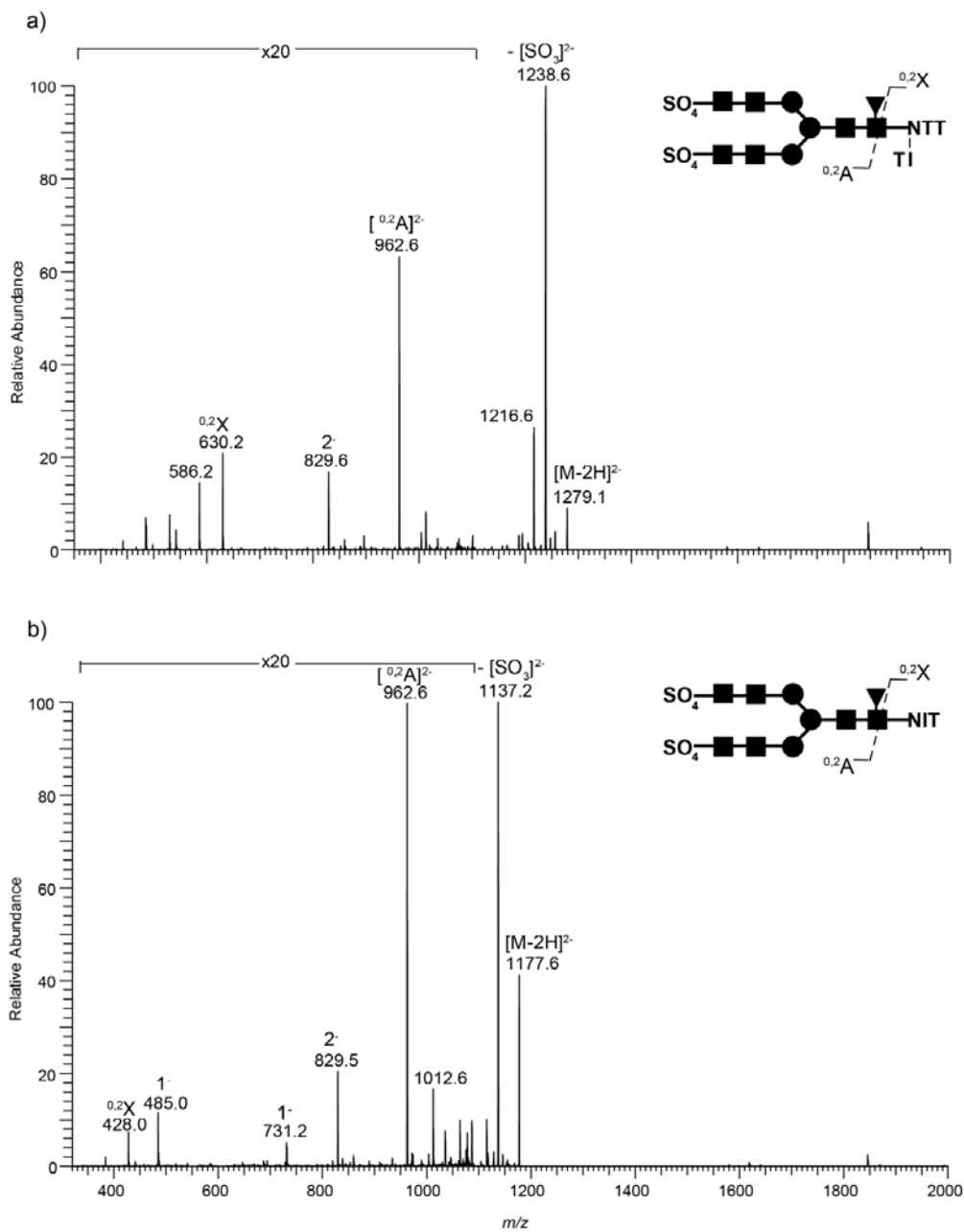


Figure 3-4: MS/MS data for doubly charged ions of sulfated glycopeptides from different glycosylation sites of eTSH. The peptide and glycan portions of these glycopeptides have been previously characterized using a combination of Edman chemistry and FTICR-MS.¹⁹ (a) MS/MS data for disulfated glycopeptide from Asn⁸² of the β -subunit. (b) MS/MS data for disulfated glycopeptide from Asn⁵⁶ glycosylation site of the α -subunit

Peptide Prediction Table			Matched Peaks		
PEPTIDE SEQUENCE	PEPTIDE MASS	$^{0,2}X$ (CHARGED)	$^{0,2}X$ ION	$^{0,2}A$ ION	IDENTIFIED PEPTIDE
MLVPKN	700.3942	782.424	630.2		TINTT
LVPKNI	682.4378	764.4676	630.3		TINTT
VPKNIT	670.4014	752.4312	630.4		TINTT
PKNITS	658.365	740.3948			
KNITSE	690.3549	772.3847			
NITSES	649.2919	731.3217			
LVPKN	569.3537	651.3835			
VPKNI	569.3537	651.3835			
PKNIT	571.333	653.3628			
KNITS	561.3123	643.3421			
NITSE	562.2599	644.2897			
VPKN	456.2696	538.2994			
PKNI	470.2853	552.3151			
KNIT	474.2803	556.3101			
NITS	433.2173	515.2471			
PKN	357.2012	439.231			
KNI	373.2326	455.2624			
NIT	346.1853	428.2151			
KN	260.1485	342.1783			
NI	245.1376	327.1674			
NIKLEN	729.4022	811.432			
IKLENH	752.4182	834.448			
KLENHT	740.3818	822.4116			
LENHTQ	740.3454	822.3752			
ENHTOC	787.2919	869.3217			
NHTOCY	821.3126	903.3424			
IKLEN	615.3593	697.3891			
KLENH	639.3341	721.3639			
LENHT	612.2868	694.3166			
ENHTQ	627.2613	709.2911			
NHTOC	658.2493	740.2791			
KLEN	502.2752	584.305			
LENH	511.2391	593.2689			
ENHT	499.2027	581.2325			
NHTQ	498.2187	580.2485			
LEN	374.1802	456.21			
ENH	398.155	480.1848			
NHT	370.1601	452.1899			
EN	261.0961	343.1259			
NH	269.1124	351.1422			
YCLTIN	782.3633	864.3931			
CLTINT	720.3477	802.3775			
LTINTT	661.3648	743.3946			
TINTTI	661.3648	743.3946			
INTTIC	720.3477	802.3775			
NTTICA	678.3007	760.3305			
CLTIN	619.3	701.3298			
LTINT	560.3171	642.3469			
TINTT	548.2807	630.3105			
INTTI	560.3171	642.3469			
NTTIC	607.2636	689.2934			
LTIN	459.2694	541.2992			
TINT	447.233	529.2628			
INTT	447.233	529.2628			
NTTI	447.233	529.2628			
TIN	346.1853	428.2151			
INT	346.1853	428.2151			
NTT	334.1489	416.1787			
IN	245.1376	327.1674			
NT	233.1012	315.131			

Peptide Prediction Table			Matched Peaks		
PEPTIDE SEQUENCE	PEPTIDE MASS	$^{0,2}X$ (CHARGED)	$^{0,2}X$ ION	$^{0,2}A$ ION	IDENTIFIED PEPTIDE
MLVPKN	700.3942	782.424	428.2	1926.6	NIT
LVPKNI	682.4378	764.4676	428.2	1926.6	TIN
VPKNIT	670.4014	752.4312	428.2	1926.6	INT
PKNITS	658.365	740.3948	428.3	1926.5	NIT
KNITSE	690.3549	772.3847	428.3	1926.5	TIN
NITSES	649.2919	731.3217	428.3	1926.5	INT
LVPKN	569.3537	651.3835	694.3	1660.5	LENHT
VPKNI	569.3537	651.3835	731.3	1623.5	NITSES
PKNIT	571.333	653.3628	731.4	1623.4	NITSES
KNITS	561.3123	643.3421			
NITSE	562.2599	644.2897			
VPKN	456.2696	538.2994			
PKNI	470.2853	552.3151			
KNIT	474.2803	556.3101			
NITS	433.2173	515.2471			
PKN	357.2012	439.231			
KNI	373.2326	455.2624			
NIT	346.1853	428.2151			
KN	260.1485	342.1783			
NI	245.1376	327.1674			
NIKLEN	729.4022	811.432			
IKLENH	752.4182	834.448			
KLENHT	740.3818	822.4116			
LENHTQ	740.3454	822.3752			
ENHTOC	787.2919	869.3217			
NHTOCY	821.3126	903.3424			
IKLEN	615.3593	697.3891			
KLENH	639.3341	721.3639			
LENHT	612.2868	694.3166			
ENHTQ	627.2613	709.2911			
NHTOC	658.2493	740.2791			
KLEN	502.2752	584.305			
LENH	511.2391	593.2689			
ENHT	499.2027	581.2325			
NHTQ	498.2187	580.2485			
LEN	374.1802	456.21			
ENH	398.155	480.1848			
NHT	370.1601	452.1899			
EN	261.0961	343.1259			
NH	269.1124	351.1422			
YCLTIN	782.3633	864.3931			
CLTINT	720.3477	802.3775			
LTINTT	661.3648	743.3946			
TINTTI	661.3648	743.3946			
INTTIC	720.3477	802.3775			
NTTICA	678.3007	760.3305			
CLTIN	619.3	701.3298			
LTINT	560.3171	642.3469			
TINTT	548.2807	630.3105			
INTTI	560.3171	642.3469			
NTTIC	607.2636	689.2934			
LTIN	459.2694	541.2992			
TINT	447.233	529.2628			
INTT	447.233	529.2628			
NTTI	447.233	529.2628			
TIN	346.1853	428.2151			
INT	346.1853	428.2151			
NTT	334.1489	416.1787			
IN	245.1376	327.1674			
NT	233.1012	315.131			

Figure 3-5: Peptide prediction table from GlycoPep ID containing the predicted and matched m/z 's for $^{0,2}X$ and/or $^{0,2}A$ ions from sulfated glycopeptides of eTSH.

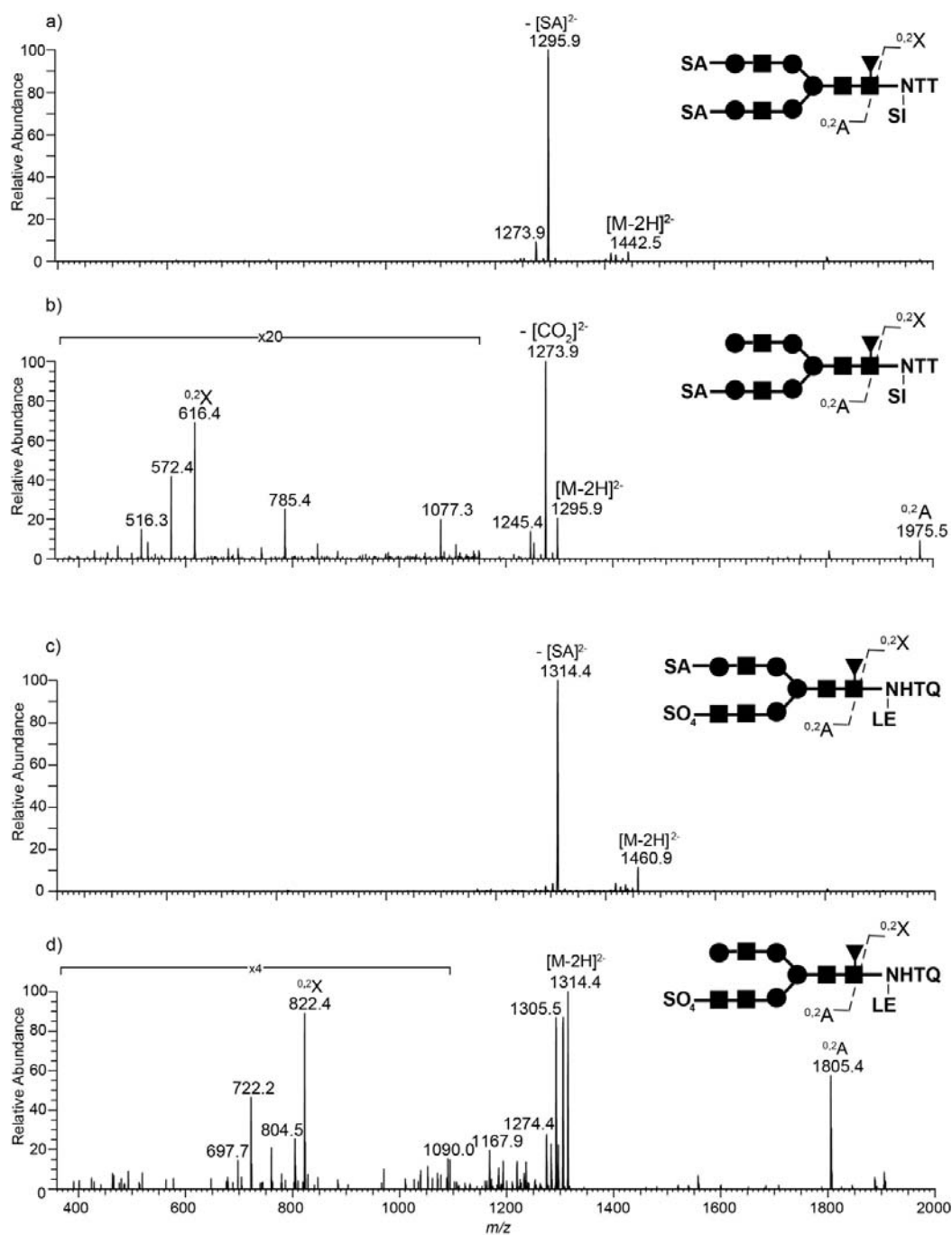


Figure 3-6: CID spectra representing doubly charged ions of glycopeptides from eFSH. The glycopeptide compositions were previously characterized using Edman chemistry and FTICR-MS.³⁹ (a) and (b) represent MS² and MS³ data respectively of an example of glycopeptides that are exclusively sialylated; (c) and (d) represent an example of MS² and MS³ data respectively, obtained from glycopeptides that are both sialylated and sulfated.

Table 3-1: Peptide identification for sulfated glycopeptides from eTSH

Parent ion <i>m/z</i>	Observed ion ^{0,2} X	Observed ion ^{0,2} A	Peptide identified	Peptide from validation data ^a
^b [861.3] ²⁻	[428.1] ⁻	[1294.1] ⁻	NIT	NIT
^b [942.3] ²⁻	[428.1] ⁻	[1456.1] ⁻	NIT	NIT
^b [1023.4] ²⁻	[428.1] ⁻	[1618.1] ⁻	NIT	NIT
^b [1064.4] ²⁻	[428.1] ⁻	[1700.1] ⁻	NIT	NIT
^c [1104.4] ²⁻	[428.1] ⁻	[889.6] ²⁻	NIT	NIT
^c [1177.4] ²⁻	[428.1] ⁻	[962.6] ²⁻	NIT	NIT
^c [1186.9] ²⁻	[593.2] ⁻	[889.7] ²⁻	LENH	LENH
^c [1227.9] ²⁻	[529.2] ⁻	[962.6] ²⁻	NTTI/TINT/INTT	INTT
^c [1278.4] ²⁻	[630.2] ⁻	[962.6] ²⁻	TINTT	TINTT

^a Data from reference 19.

^b Previously characterized mono-sulfated, glycopeptides.

^c Previously characterized di-sulfated, glycopeptides.

3.3.3 Application to complex negatively charged glycopeptides from eFSH

Having demonstrated the efficacy of using GlycoPep ID in identifying the peptide portion of sulfated glycopeptides from (-) ESI-MS/MS data, this method is extended to identify the peptide moieties from more complex glycopeptides, such as those from follicle stimulating hormone (FSH). FSH belongs to the same pituitary glycoprotein hormone family as TSH. One main difference between FSH and TSH is the fact that unlike TSH, FSH contains four glycosylation sites, and glycopeptides released from this hormone contain glycans that are exclusively sialylated and those that are sulfated and sialylated. The FSH sample was prepared in the same manner as the TSH sample, and the data was analyzed using the same approach as well. The amino acid sequence for FSH is readily available from Swiss-Prot, and is shown in Figure 3-2. Two example analyses of this hormone using GlycoPep ID are described below. These examples are for a glycopeptide that is exclusively sialylated and one that is both sialylated and sulfated.

3.3.4 CID experiment data for glycopeptides containing sialic acid only

MS/MS experiments on sialylated glycopeptides indicated that loss of sialic acid is very favorable, and when this occurs, virtually no other product ions are present in the spectrum. This is consistent with previous

researchers' findings who have documented the extreme lability of sialic acid.^{13, 40} Despite the lability of sialic acid, it does not represent a significant obstacle for our analysis, since MS³ experiments can be performed on the product ion that corresponds to the loss of sialic acid, and the product ions from that experiment can be used to identify the peptide.

Figure 3-6a and b represent an example of MS² and MS³ data obtained from activation of a doubly charged glycopeptide containing two sialic acids. As shown in Figure 3-6a, MS/MS data of this glycopeptide at *m/z* 1442 provides no informative product ions. The base peak, *m/z* 1296, results from loss of sialic acid. An MS³ experiment performed on this ion is shown in Figure 3-6b and the peak list from MS³ data was input in GlycoPep ID. The following input parameters were used: the eFSH protein sequence, proteinase K for enzyme, charge state of -1, mass tolerance of ±0.1 Da, carbamidomethyl for cysteine modification, and calculation for ^{0,2}A ion was not selected. GlycoPep ID generated only one product ion, *m/z* 616, from the potential ^{0,2}X ions in the prediction table within a mass error of 0.1 Da (Figure 3-7A). This product ion corresponds to the peptide sequence, SINTT from the prediction table. For the glycopeptide depicted in Figures 3-6a and b, the complementary ^{0,2}A ion is observed in the mass spectrum at *m/z* 1975 in Figure 3-6b; therefore SIN²⁴TT is definitively identified as the peptide moiety in this glycopeptide. This peptide sequence is consistent with that identified from the high-resolution FTICR-MS data and Edman data.³⁹ The

final results for this glycopeptide, along with others from FSH, are summarized in Table 3-2. Again, in each case the identified peptide matched the validation data. For glycopeptides at m/z 1376, 1449, 1500, 1543, and m/z 1573, the $^{0,2}A$ ions are not observed, because they are outside the upper mass limit of the instrument used in this analysis. This does not affect the overall results for this study, since each of these disialylated glycopeptides only had one product ion that matched the predicted masses of $^{0,2}X$ ions in prediction table in Figure 3-7A. Thus, the presence of their complementary $^{0,2}A$ ions is not necessary for the correct peptide to be identified. However, had there been more than one product ion that matched the predicted masses of $^{0,2}X$ ions in the prediction table, the presence of the complementary $^{0,2}A$ ion would be necessary to verify the correct peptide moiety.

3.3.5 CID experiments for glycopeptides containing both sulfate and sialic acid

Figure 3-6c and d represent CID data obtained from a doubly charged ion of a glycopeptide from eFSH that contains both sialic acid and sulfate. As in the case of glycopeptides containing sialic acid only, the MS/MS data of glycopeptides containing both sialic acid and sulfate provide no useful information for identifying the peptide present. Since sialic acid is the most labile group in the molecule, loss of sialic acid is the only product ion

observed during MS/MS experiments, at m/z 1314 (Figure 3-6c). When MS³ experiments are performed on this ion, several product ions are observed that could correspond to the peptide moiety for this glycopeptide; see Figure 3-6d. Analysis of this data using GlycoPep ID generated three plausible matches: LENHTQ, KLENHT, and SCELTN. Specifically, m/z 822 could correspond to the peptides LEN⁸²HTQ or KLEN⁸²HT and m/z 804 correspond to the peptide SCELTN. The abundance of m/z 804 is ~28% while that of m/z 822, which is the most abundant product ion in the mass range of interest, is 100%. Using the same approach as described above, a search for each complementary ^{0,2}A ion was performed on the CID data so as to identify the correct peptide moiety. A close inspection of the mass spectrum in Fig. 6-3d reveals that only m/z 822 has its singly charged complementary ion present at m/z 1805. This information indicates that either LEN⁸²HTQ or KLEN⁸²HT is the peptide present in this glycopeptide. Since LEN⁸²HTQ and KLEN⁸²HT are isobaric compounds, they can be discriminated by the high-resolution data from FTICR-MS, which confirms LEN⁸²HTQ is the correct peptide³⁹ as indicated by Table 3-2. Even though the low-resolution method could not distinguish between LENHTQ and KLENHT, both these glycopeptides map to the same glycosylation site, so distinguishing between the two sequences is not necessary.

CID data from all the other glycopeptides that are both sialylated and sulfated produce similar information. A summary of all the glycopeptides

from FSH investigated in this study is shown in Table 3-2. All the identified peptide moieties from GlycoPep ID match the assignments made when these samples were previously analyzed, using a combination of Edman sequencing and FTICR-MS analysis.³⁹

3.4 Conclusion

The results from this study clearly demonstrate that the peptide moieties of negatively charged glycopeptides can be identified by matching product ions in CID data to a prediction table generated in an automated fashion, from the web-based program, GlycoPep ID. The versatility of the method using a non-specific protease, proteinase K and the automated data analysis of glycosylation in a site specific manner was demonstrated by identifying the peptide moieties of glycopeptides from two different glycoprotein hormones that were exclusively sialylated or sulfated, or were both sialylated and sulfated. All 27 peptide moieties in this study were correctly identified by GlycoPep ID and validated using data from a combination of Edman chemistry and high resolution FTICR-MS analysis. This technique represents an important advance in glycosylation profiling because it solves one of the most difficult problems of using proteinase K in glycopeptide analysis: Determining where the enzyme cleaved the protein.

Peptide Prediction Table				Matched Peaks	
PEPTIDE SEQUENCE	PEPTIDE MASS	^{0,2} X (CHARGED)	^{0,2} X ION	^{0,2} A ION	IDENTIFIED PEPTIDE
MLVPKN	700.3942	782.424	616.3	2010.7	SINTT
LVPKNI	682.4378	764.4676	616.4	2010.6	SINTT
VPKNIT	670.4014	752.4312			
PKNITS	658.365	740.3948			
KNTSE	690.3549	772.3847			
NITSES	649.2919	731.3217			
LVPKN	569.3537	651.3835			
VPKNI	569.3537	651.3835			
PKNIT	571.333	653.3628			
KNITS	561.3123	643.3421			
NITSE	562.2599	644.2897			
VPKN	456.2696	538.2994			
PKNI	470.2853	552.3151			
KNIT	474.2803	556.3101			
NITS	433.2173	515.2471			
PKN	357.2012	439.231			
KN	373.2326	455.2624			
NIT	346.1853	428.2151			
KN	260.1485	342.1783			
NI	245.1376	327.1674			
NKLEN	729.4022	811.432			
IKLENH	752.4182	834.448			
KLENHT	740.3818	822.4116			
LENHTO	740.3454	822.3752			
ENHTOC	787.2919	869.3217			
NHTOCY	821.3126	903.3424			
IKLEN	615.3593	697.3891			
KLENH	639.3341	721.3639			
LENHT	612.2868	694.3166			
ENHTO	627.2613	709.2911			
NHTOC	658.2493	740.2791			
KLEN	502.2752	584.305			
LENH	511.2391	593.2689			
ENHT	499.2027	581.2325			
NHTO	498.2187	580.2485			
LEN	374.1802	456.21			
ENH	398.155	480.1848			
NHT	370.1601	452.1899			
EN	261.0961	343.1259			
NH	269.1124	351.1422			
SCELTN	722.2905	804.3203			
CELTNI	748.3426	830.3724			
ELTNT	689.3597	771.3895			
LTNIT	673.4012	755.431			
TNITIA	631.3542	713.384			
NITIAV	629.3749	711.4047			
CELTN	635.2585	717.2883			
ELTN	588.312	670.3418			
LTNIT	560.3171	642.3469			
TNIT	560.3171	642.3469			
NITIA	530.3065	612.3363			
ELTN	475.2279	557.2577			
LTN	459.2694	541.2992			
TNIT	447.233	529.2628			
NITI	459.2694	541.2992			
LTN	346.1853	428.2151			
TNI	346.1853	428.2151			
NIT	346.1853	428.2151			
TN	233.1012	315.131			
NI	245.1376	327.1674			
FCISIN	752.3527	834.3825			
CISINT	706.332	788.3618			
ISINTT	647.3491	729.3789			
SINTTW	720.3443	802.3741			
INTTWC	793.3429	875.3727			
NTTWCA	751.2959	833.3257			
CISIN	605.2843	687.3141			
ISINT	546.3014	628.3312			
SINTT	534.265	616.2948			
INTTW	633.3123	715.3421			
NTTWC	680.2588	762.2886			
ISIN	445.2537	527.2835			
SINT	433.2173	515.2471			
INTT	447.233	529.2628			
NTTW	520.2282	602.258			
SIN	332.1696	414.1994			
INT	346.1853	428.2151			
NTT	334.1489	416.1787			
IN	245.1376	327.1674			
NT	233.1012	315.131			

Peptide Prediction Table				Matched Peaks	
PEPTIDE SEQUENCE	PEPTIDE MASS	^{0,2} X (CHARGED)	^{0,2} X ION	^{0,2} A ION	IDENTIFIED PEPTIDE
MLVPKN	700.3942	782.424			
LVPKNI	682.4378	764.4676	804.3	1823.5	SCELTN
VPKNIT	670.4014	752.4312	822.4	1805.4	KLENHT
PKNITS	658.365	740.3948	822.4	1805.4	LENHTO
KNTSE	690.3549	772.3847	822.5	1805.3	KLENHT
NITSES	649.2919	731.3217	822.5	1805.3	LENHTO
LVPKN	569.3537	651.3835			
VPKNI	569.3537	651.3835			
PKNIT	571.333	653.3628			
KNITS	561.3123	643.3421			
NITSE	562.2599	644.2897			
VPKN	456.2696	538.2994			
PKNI	470.2853	552.3151			
KNIT	474.2803	556.3101			
NITS	433.2173	515.2471			
PKN	357.2012	439.231			
KN	373.2326	455.2624			
NIT	346.1853	428.2151			
KN	260.1485	342.1783			
NI	245.1376	327.1674			
NKLEN	729.4022	811.432			
IKLENH	752.4182	834.448			
KLENHT	740.3818	822.4116			
LENHTO	740.3454	822.3752			
ENHTOC	787.2919	869.3217			
NHTOCY	821.3126	903.3424			
IKLEN	615.3593	697.3891			
KLENH	639.3341	721.3639			
LENHT	612.2868	694.3166			
ENHTO	627.2613	709.2911			
NHTOC	658.2493	740.2791			
KLEN	502.2752	584.305			
LENH	511.2391	593.2689			
ENHT	499.2027	581.2325			
NHTO	498.2187	580.2485			
LEN	374.1802	456.21			
ENH	398.155	480.1848			
NHT	370.1601	452.1899			
EN	261.0961	343.1259			
NH	269.1124	351.1422			
SCELTN	722.2905	804.3203			
CELTNI	748.3426	830.3724			
ELTNT	689.3597	771.3895			
LTNIT	673.4012	755.431			
TNITIA	631.3542	713.384			
NITIAV	629.3749	711.4047			
CELTN	635.2585	717.2883			
ELTN	588.312	670.3418			
LTNIT	560.3171	642.3469			
TNIT	560.3171	642.3469			
NITIA	530.3065	612.3363			
ELTN	475.2279	557.2577			
LTN	459.2694	541.2992			
TNIT	447.233	529.2628			
NITI	459.2694	541.2992			
LTN	346.1853	428.2151			
TNI	346.1853	428.2151			
NIT	346.1853	428.2151			
TN	233.1012	315.131			
NI	245.1376	327.1674			
FCISIN	752.3527	834.3825			
CISINT	706.332	788.3618			
ISINTT	647.3491	729.3789			
SINTTW	720.3443	802.3741			
INTTWC	793.3429	875.3727			
NTTWCA	751.2959	833.3257			
CISIN	605.2843	687.3141			
ISINT	546.3014	628.3312			
SINTT	534.265	616.2948			
INTTW	633.3123	715.3421			
NTTWC	680.2588	762.2886			
ISIN	445.2537	527.2835			
SINT	433.2173	515.2471			
INTT	447.233	529.2628			
NTTW	520.2282	602.258			
SIN	332.1696	414.1994			
INT	346.1853	428.2151			
NTT	334.1489	416.1787			
IN	245.1376	327.1674			
NT	233.1012	315.131			

Figure 3-7: Peptide prediction table from GlycoPep ID containing the predicted and matched m/z 's for ^{0,2}X and/or ^{0,2}A ions from sulfated glycopeptides of eFSH.

Table 3-2: Summary of all analyzed glycopeptides from eFSH using GlycoPep ID.

<i>m/z</i> for parent ion (2H ⁺)	Type of modification	<i>m/z</i> for ^{0,2} X observed	<i>m/z</i> for ^{0,2} A observed	Peptide identified	Peptide from validation data [#]
1043*	Sulfation	428	1659	N ^{7/56} IT/LTN ⁷	N ^{7/56} IT
1104*	Sulfation	428	1700	N ^{7/56} IT/LTN ⁷	N ^{7/56} IT
1189	Sialylation and sulfation	428	1659	N ^{7/56} IT/LTN ⁷	N ^{7/56} IT
1227*	Sulfation	529	[962.82] ²⁻	IN ²⁴ TT	IN ²⁴ TT
1262	Sialylation and sulfation	428	1805	N ^{7/56} IT/LTN ⁷	N ^{7/56} IT
1301*	Sulfation	822	1700	LEN ⁸² HTQ/ KLEN ⁸² HT	LEN ⁸² HTQ
1313	Sialylation and sulfation	529	1805	IN ²⁴ TT	IN ²⁴ TT
1347	Sialylation	428	1975	N ^{7/56} IT/LTN ⁷	N ^{7/56} IT
1356	Sialylation and sulfation	616	1805	SIN ²⁴ TT	SIN ²⁴ TT
1376	Sialylation	428	Out of range	N ^{7/56} IT/LTN ⁷	N ^{7/56} IT
1386	Sialylation and sulfation	822	1659	LEN ⁸² HTQ/ KLEN ⁸² HT	LEN ⁸² HTQ
1398	Sialylation	529	1975	IN ²⁴ TT	IN ²⁴ TT
1442	Sialylation	616	1976	SIN ²⁴ TT	SIN ²⁴ TT
1449	Sialylation	428	Out of range	N ^{7/56} IT/LTN ⁷	N ^{7/56} IT
1460	Sialylation and sulfation	822	1805	LEN ⁸² HTQ/ KLEN ⁸² HT	LEN ⁸² HTQ
1500	Sialylation	529	Out of range	IN ²⁴ TT	IN ²⁴ TT
1543	Sialylation	616	Out of range	SIN ²⁴ TT	SIN ²⁴ TT
1573	Sialylation	822	Out of range	LEN ⁸² HTQ/ KLEN ⁸² HT	LEN ⁸² HTQ

***Ions that provided peptide information in MS/MS. Peptide information for the rest of the ions was obtained during MS³ experiments.**

Data from reference 39

3.5 References

1. Zaia, J. *Mass Spectrom Rev.* **2004**, 23,161-227.
2. Blom, N.; Sicheritz-Ponten, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. *Proteomics* **2004**, 4,1633-1649.
3. Sullivan, B.; Addona, T. A.; Carr, S. A. *Anal. Chem.* **2004**, 76, 3112-3118.
4. Zhen, Y.; Caprioli, R.; Staros, J. *Biochemistry* **2003**, 42, 5478-5492.
5. Imre, T.; Schlosser, G.; Pocsfalvi, G.; Siciliano, R.; Molnar-Szollosi, E.; Kremmer, T; Malorni, A.; Vekey, K. *J. Mass Spectrom.* **2005**, 40, 1472-1483.
6. Yeh, J.; Seals, J. R.; Murphy, C. I.; van Halbeek, H.; Cummings, R. D. *Biochemistry* **1993**, 32, 11087-99.
7. Cutalo, J. M.; Deterding, L. J.; Tomer, K. B. *J. Am. Soc. Mass Spectrom.* **2004**, 15, 1545-1555.
8. Zhu, X.; Borchers, C.; Bienstock, R. J.; Tomer, K. B. *Biochemistry* **2000**, 39, 11194-11204.
9. Carr, S. A.; Huddleston, M. J.; Bean, M. F. *Protein Sci.* **1993**, 2, 183-96.
10. Diop, N. K.; Hrycyna, C. A. *Biochemistry* **2005**, 44, 5420-5429.
11. Freeze, H. H.; Aebi, M. *Curr. Opin. Struct. Biol.* **2005**, 15, 490-498.
12. Harvey, D. J. *Expert Rev Proteomics.* **2005**, 2, 87-101.
13. Papac, D. I.; Wong, A.; Jones, A. J. S. *Anal. Chem.* **1996**, 68, 3215-3223.
14. Ethier, M.; Krokhin, O.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Perreault, H. *Rapid Commun. Mass Spectrom.* **2005**, 19, 721-727.
15. Harazono, A.; Kawasaki, N.; Itoh, S.; Hashii, N.; Ishii-Watabe, A.; Kawanishi, T.; Hayakawa, T. *Anal. Biochem.* **2006**, 348, 259-268.
16. Harmon, B. J.; Gu, X.; Wang, I. C. *Anal. Chem.* **1996**, 68,1465-73.

17. Jiang, H.; Irungu, J.; Desaire, H. *J. Am. Soc. Mass Spectrom.* **2005**, 16, 340-348.
18. Jiang, H.; Desaire, H.; Butnev, V. Y.; Bousfield, George R. *J. Am. Soc. Mass Spectrom.* **2004**, 15, 750-758.
19. Irungu, J.; Dalpathado, D. S.; Go, E. P.; Jiang, H.; Ha, Hy-Vy; Bousfield, G. R.; Desaire, H. *Anal. Chem.* **2006**, 78, 1181-1190.
20. An, H. J.; Peavy, T. R.; Hedrick, J. L.; Lebrilla, C. B. *Anal. Chem.* **2003**, 75, 5628-5637.
21. Huddleston, M. J.; Bean, M. F.; Carr, S. A. *Anal. Chem.* **1993**, 65, 877-84.
22. Nemeth, J. F.; Hochensang, G. P.; Marnett, L. J.; Caprioli, R. M. *Biochemistry* **2001**, 40, 3109-3116.
23. Kvaratskhelia, M.; Clark, P. K.; Hess, S.; Melder, D. C.; Federspiel, M. J.; Hughes, S. H. *Virology* **2004**, 326, 171-181.
24. Yang, Y.; Orlando, R. *Rapid Commun. Mass Spectrom.* **1996**, 10, 932-936.
25. Juhasz, P.; Martin, S. A. *Int. J. Mass Spectrom. Ion Process.* **1997**, 169/170 217-230.
26. Stephens, E.; Sugars, J.; Maslen, S. L.; Williams, D. H.; Packman, L. C.; Ellar, D.J. *Eur. J. Biochem.* **2004**, 271, 4241-4258.
27. von Witzendorff, D.; Ekhlasi-Hundrieser, M.; Dostalova, Z.; Resch, M.; Rath, D.; Michelmann, H.; Toepfer-Petersen, E. *Glycobiology* **2005**, 15, 475-488.
28. Kamar, M.; Alvarez-Manilla, G.; Abney, T.; Azadi, P.; Kolli, V. S. K.; Orlando, R.; Pierce, M. *Glycobiology* **2004**, 14, 583-592.
29. Krokhin, O.; Ens, W.; Standing, K. G.; Wilkins, J.; Perreault, H. *Rapid Commun. Mass Spectrom.* **2004**, 18, 2020-2030.
30. Haegglund, P.; Bunkenborg, J.; Elortza, F.; Jensen, O. N.; Roepstorff, P. *J. Proteome Res.* **2004**, 3, 556-566.

31. Harazono, A.; Kawasaki, N.; Kawanishi, T.; Hayakawa, T. *Glycobiology* **2005**, 15, 447-62.
32. Conboy, J. J.; Henion, J. D. *J. Am. Soc. Mass Spectrom.* **1992**, 3, 804-14.
33. Bykova, N. V.; Rampitsch, C.; Krokhin, O.; Standing, K. G.; Ens, W. *Anal. Chem.* **2006**, 78, 1093-1103.
34. Rahbek-Nielsen, H.; Roepstorff, P.; Reischl, H.; Wozny, M.; Koll, H.; Haselbeck, A. *J. Mass Spectrom.* **1997**, 32, 948-958.
35. Wuhrer, M.; Hokke, C. H.; Deelder, A. M. *Rapid Commun. Mass Spectrom.* **2004**, 18, 1741-1748.
36. Gooley, A. A.; Classon, B. J.; Marschalek, R.; Williams, K. L. *Biochem. Biophys. Res. Commun.* **1991**, 178, 1194-201.
37. Zeng, R.; Xu, Q.; Shao, Xiao-Xia; Wang, Ke-Yi; Xia, Qi-Chang. *Eur. J. Biochem.* **1999**, 266, 352-358.
38. Larsen, M. R.; Hojrup, P.; Roepstorff, P. *Mol. Cell Proteomics* **2005**, 4, 107-119.
39. Dalpathado, D.S.; Irungu, J.; Go, E. P.; Butnev, V.Y.; Norton, K.; Bousfield, G. R.; Desaire, H. *Biochemistry*, **2006**, 45, 8665-8673.
40. Froesch, M.; Bindila, L.; Zamfir, A.; Peter-katalinic, J. *Rapid Commun. Mass Spectrom.* **2003**, 17, 2822-2832.
41. Hakansson, K.; Cooper, H. J.; Emmett, M. R.; Costello, C. E.; Marshall, A. G.; Nilsson, C.L. *Anal. Chem.* **2001**, 73, 4530-4536.
42. Bousfield, G. R.; Ward, D. N. *J. Bio. Chem.* **1984**, 259, 1911-21.

CHAPTER 4

Comparison of LC/ESI-FTICR MS vs MALDI-TOF/TOF MS for glycopeptide analysis of a highly glycosylated protein: HIV Envelope glycoprotein

4.1 Introduction

Glycoproteomics is a newly emerging field that involves the characterization of protein glycosylation. It is widely accepted that glycosylation is by far the most common post translational modification present in both eukaryotic and prokaryotic proteins.^{1, 2} This modification plays a major role in proteins' biological and cellular processes, and it influences their physiochemical properties.³⁻⁵ Glycans have also been shown to play a vital role in various parasitic, bacterial and viral disease infections.⁶ For instance, interactions and fusion of the human immunodeficiency virus (HIV) with its target host cells is mediated by its envelope protein, gp160, which is extensively glycosylated protein with over 50% of its mass comprising of glycans.⁷⁻¹⁰ The high population and diverse range of glycan structures on this protein act as a shield for the virus against the immune system by masking epitopes that could be targeted for immune attack.^{8, 11-17} Consequently, defining the structures and locations of glycans in the HIV envelope protein is important in understanding how variation in glycosylation affects the protein's function, and in this particular example, glycosylation information may also provide valuable structural insight into current HIV vaccine candidates, which is useful in vaccine development. (Ref eden's

paper) . To acquire this information, sensitive, rapid, and reliable methods for mapping and profiling glycosylation in proteins are required.

Unlike in the proteomics and glycomics fields where methods of analysis are well established, analytical methods of analysis in the glycoproteomics field are still underway. Mass spectrometry (MS) has gained a widespread use in protein glycosylation analysis and has become an indispensable, powerful analytical technique in the field of glycoproteomics, due to its high sensitivity and selectivity. Analysis of protein glycosylation by mass spectrometry is typically achieved by two main approaches: Either glycans can be released from the peptide backbone enzymatically or chemically, or the glycoprotein can be subjected to a protease digestion, producing a mixture of peptides and glycopeptides. The latter approach is advantageous to releasing glycans from the protein, since it does not require extra sample manipulation. and it allows for site-specific glycosylation profiling.¹⁸ However, there are several obstacles encountered when using a glycopeptide-based MS analysis. For example, glycopeptides exhibit poor ionization efficiency and their signal is usually suppressed by non-glycosylated peptides. In addition, most glycosylation sites contain various glycoforms, and each glycoform may exist at low concentration in the total glycopeptide mixture.^{19, 20} To obviate these obstacles, it is often necessary to perform an enrichment or chromatographic separation prior to MS analysis. Several studies have addressed this issue and proposed effective

enrichment or/and chromatographic methods that can be utilized prior to MS analysis of glycopeptides.^{3, 20}

Although glycopeptide-based MS approaches are typically used for glycoprotein analysis, so far there is no consensus as to which MS approach would provide the most glycosylation information, especially for a complex glycoprotein. Recent advances in glycopeptide-based MS analysis have been achieved by two emerging platforms, online LC/ESI-FTICR-MS and offline HPLC/ MALDI-TOF/TOF MS. These methods are known for their unique high resolution and high mass accuracy capabilities, along with their ability to accommodate MS/MS experiments. MALDI-TOF/TOF is widely used partly because it has a higher dynamic range and has a high tolerance to salts and other contaminants. Besides, the complexity of data obtained in ESI-FTICR-MS due to the presence of multiply charged ions and formation of salt adducts greatly complicates data interpretation of heterogeneous glycopeptide mixtures.²¹ However, unlike offline HPLC/MALDI-MS, online LC-ESI-FTICR-MS efficiently provides great deal of information in a single experiment.²² Furthermore, glycan-specific ions can be selectively identified from full MS¹ scan and used to trigger subsequent MSⁿ scans during chromatographic separation of complex digest mixture, thereby providing a plethora of information about the glycopeptides in question.²³ On the contrary, MSⁿ experiments cannot be performed by MALDI-TOF/TOF MS, which limits the amount of information that can be acquired using this

platform. In addition, MALDI analyses suffer from matrix-dependent ionization and fragmentation processes.²⁴⁻²⁷ The type of matrix used for glycopeptide analysis largely influences the extent and type of fragmentation ions produced during MALDI-MS/MS experiments.²⁶ Since neither the MALDI nor ESI platforms stand out as a clearly superior approach, we performed a head-to-head comparison on both platforms, using a highly complex glycoprotein sample, in order to investigate the merits and limitations of each method.

Herein, we present a detailed study to investigate the merits of offline HPLC/MALDI-TOF/TOF and online LC-ESI-FTICR when used to provide glycosylation information of a recently characterized glycoprotein containing 31 potential glycosylation sites.²⁸ Specifically, we employed the two platforms to analyze the number of glycopeptides and quality of MS data obtained during the analysis of the glycoprotein, CON-S gp140 Δ CFI, a synthetic form of the envelope protein found on the HIV virus (gp160).²⁹ To ensure that the intrinsic worth of each platform was fully exploited, we determined how well each platform could answer several specific research questions that will eventually contribute in understanding how glycosylation affects the function and immunogenicity of the *env* protein. These questions included: how many of the 31 potential glycosylation sites, if glycosylated, could be detected by each technique; what is the extent of glycosylation coverage provided by each platform, for each glycosylation site; what type of confirmatory

information can be obtained on both the peptide and glycan portions of the glycopeptides identified using collision induced dissociation (CID) experiments. Our results revealed significant differences in the glycosylation sites detected by using each method, the population of glycoforms identified and the type of structural information obtained on either the peptide or glycan portion of the identified glycopeptides. These results suggest that the two techniques are highly complementary, and when possible, the glycosylation information is maximized by combining the two platforms.

4.2 EXPERIMENTAL SECTION

4.2.1 Materials and Reagents

Purified CON-S gp140 Δ CFI protein was produced as recombinant vaccinia virus expressed protein from Duke Human Vaccine Research Institute in Durham, as described previously.²⁹ Urea, ethylenediaminetetraacetic acid (EDTA), dithiothreitol (DTT), iodoacetamide (IAA), HPLC grade acetonitrile (ACN), ammonium bicarbonate, trizma hydrochloride and base, formic acid, 2,5-dihydroxybenzoic acid (DHB), and α -cyano-4-hydroxycinnamic acid (CHCA), were all purchased from Sigma-Aldrich. Proteomics grade trypsin was obtained from Promega (Madison, WI). N-Glycosidase F (PNGase F) from *Elizabethkingia meningosepticum* was obtained from CalBioChem (San Diego, CA). Water used for these

studies was purified using a Millipore Direct-Q3 Water Purification System (Billerica, MA).

4.2.2 Trypsin digestion of CON-S gp140 Δ CFI protein

Approximately 300 μ g of protein was prepared in 100 mM Tris-HCl buffer, containing 6M urea and 3mM EDTA, pH 7.5. The protein was reduced for 1 hour with 15 mM DTT and alkylated for another hour at room temperature with 50 mM IAA. The excess IAA was neutralized by adding DTT, to a final concentration of 40 mM. Extra buffer solution was added to reduce the concentration of Urea to about 2M. Trypsin was added at a protein:enzyme ratio of 30:1 (w/w) to generate glycopeptides. The protein solution was incubated overnight at 37 °C. The reaction was quenched the following day by adding 1 μ L of concentrated acetic acid. Two aliquots were removed from the total digest, and each aliquot was subjected to either online LC/ESI-FTICR or offline HPLC fractionation, prior to MALDI-TOF/TOF.

4.2.3 Reverse phase HPLC fractionation

The tryptic glycopeptides/peptides mixture was purified and separated on a Shimadzu model HPLC system. For each run, 20 μ L of the tryptic digest was injected onto a C18 column (150 \times 4.6mm, 5 μ M, Alltech, Deerfield, IL) at a flow rate of 1mL/min. Purified water and HPLC grade ACN each containing 0.1% formic acid were used as mobile phase A and B respectively, with a linear gradient from 5% to 40% B over 50 min, followed by a ramp to 95% B in 10 min.³⁰ Fractions were manually collected every 1 min for 60 min. Each

fraction was evaporated to dryness on a CentriVap (Labconco Corporation, KC, MO) before reconstituting with 10 μ L of water. The reconstituted fractions were first screened and analyzed by MALDI-TOF/TOF and all fractions containing glycopeptides were then deglycosylated and reanalyzed by MALDI-TOF/TOF.

4.2.4 Deglycosylation

Reconstituted glycopeptide fractions were enzymatically deglycosylated using PNGase F (CalBioChem) by applying the protocol recommended by the manufacturer. Briefly, each enriched glycopeptide fraction was deglycosylated by adding 4 μ L of PNGase F and 25 μ L of 20 mM NH_4HCO_3 (pH = 8), and then incubated overnight at 37 °C. The reaction was stopped by boiling and analyzed by MALDI-MS.

4.2.5 MALDI-TOF/TOF MS analysis

A combination of DHB and CHCA (1:1 V/V) matrixes was used and mixed with each sample (1:1 by volume). Approximately 0.75 μ L of the mixture was spotted on a stainless steel MALDI target plate (Applied Biosystems, Foster City, CA) and air-dried. All MALDI MS and MS/MS data was acquired in the reflectron mode on an Applied Biosystems 4700 proteomics analyzer mass spectrometer. The samples were irradiated by a 355 nm Nd-YAG laser (355 nm) at 200 Hz. The acceleration voltage was 25 kV. Each mass spectrum was generated by averaging 3200 laser shots. The laser intensity was optimized to give the best signal-to-noise ratio and

resolution for each sample. All the data were processed in Data Explorer version 4.5 (Applied Biosystems). Glycopeptide analysis was performed by using the high-resolution MALDI-TOF/TOF MS data in conjunction with our previously described web-based tool, (GlycoPep DB),³¹ to assign glycopeptide compositions. The assigned compositions were then confirmed by MALDI-MS/MS experiments.

4.2.6 Capillary LC/ESI- FTICR MS analysis

Analysis of the tryptic glycopeptides on LC/ESI-FTICR-MS was performed by using a Dionex Ultimate capillary LC system (Sunnyvale, CA) equipped with a FAMOUS well plate autosampler coupled to a high-resolution Thermo Finnigan linear ion trap-Fourier Transform Ion Cyclotron Resonance mass spectrometer, LTQ-FTICR-MS, (San Jose, CA) equipped with a 7 Telsa actively shielded magnet. Samples were loaded onto a Famous well plate autosampler and 5 μ L of the tryptic digest was injected onto an LC Packings C18 PeMapTM 300 column (300 μ m i.d \times 15cm, 5 μ m, 300 Å). Water and HPLC grade ACN, each containing 0.06% formic acid, were used as mobile phase A and B respectively, with a linear gradient starting from 5% to 40% B over 50 min, followed by a ramp to 95% B in 10 min. The eluting solution was directly infused into the mass spectrometer at a flow rate of 5 μ L/min.

High resolution data was acquired on the FTICR MS by maintaining resolution at 50,000, for m/z 400. The instrument was externally calibrated

prior to the analysis, over the entire mass range of interest. The data was acquired in the mass range of m/z 800-2000 using a spray voltage of approximately 4.0 kV. N_2 was used as a nebulizing gas at 20 psi, and the capillary temperature was maintained between 200-230 °C. Data was acquired and processed using Xcalibur 1.4 SR1 software (Thermo Finnigan San Jose, CA). The glycopeptide compositions were assigned using the high resolution data together with GlycoPep DB as described previously.³¹

4.2.7 CID Experiments in LC/ESI-FTICR MS

All MS/MS data was acquired in the linear ion trap of the hybrid LTQ-FTICR in a data-dependent scanning fashion. Data dependent MS/MS data was acquired for the three most intense ions observed in full MS^1 scan, using a dynamic exclusion window. To maximize the number of data dependent MS/MS data collected for the glycopeptides ions observed in full MS^1 scan data, three more scan events were set with each subsequent scan event selecting the 2nd, 3rd and 4th three most intense ions from MS^1 data. If a neutral loss of a hexose or a HexNAc was detected in these scans, an MS^3 scan event was triggered. Each selected precursor ion was activated for 30 ms with q_z value of 0.25 and an isolation width of 3 Da. Activation amplitudes were in the range of 22-25% as defined by the instrument software.

4.2.8 Data analysis

To interpret the high-resolution data acquired from LC/ESI-FTICR MS, several steps were undertaken. The first step was to determine if the peaks observed in MS¹ were glycopeptides or not. In order to verify this, the lower mass range region of MS/MS spectra of those peaks were examined for the presence of glycan characteristic product ions like m/z 528 [HexNAc+2Hex+H]⁺, m/z 690 [HexNAc+3Hex+H]⁺, m/z 893 (triamannosyl chitibose core), or m/z 657 [HexNAc+Hex+Sialic Acid+H]⁺. If any of these ions was observed, the next step was to input the MS/MS peaklist of the glycopeptides in question into our newly developed web-based tool, GlycoPep ID. A complete description of how this tool operates was provided previously.³² Briefly, GlycoPep ID uses characteristic fragment ions, such as ^{0,2}X ion [Peptide+83-H]⁻ or Y₁ ion [Peptide+203+H]⁺, observed in MS/MS to predict the potential peptide portion of the glycopeptides in question. From LC/ESI-MS/MS data in the positive ion mode, each of the scan events provided a specific characteristic fragmentation ion, Y₁, a glycosidic bond cleavage that occurs at the inner core of *N*-acetyl glucosamine (HexNAc) attached to the peptide, and this ion was automatically predicted by GlycoPep ID, thus identifying the peptide portion of the glycopeptides in question. The identified peptide portion was then inputted into GlycoPep DB, described previously³¹, which utilizes the high resolution MS¹ peaklist to generate all the plausible glycan compositions attached to that specific

peptide. All the glycopeptide compositions outputted were then inspected manually and verified by using MS¹ and MS/MS data. MS/MS data was also used to confirm the assigned glycopeptide compositions and to obtain composition information about the glycan portions.

For MALDI-TOF/TOF data analysis, MALDI-MS/MS data obtained from each glycopeptide fraction was first analyzed to identify the ^{0,2}X ion [Peptide+83+H]⁺, a characteristic product ion that is typically observed in MALDI-MS/MS data of glycopeptides. This ion corresponds to the peptide portion attached to a portion of the glycan, which remains attached to the peptide after the cross ring cleavage.³³ The identified peptide portion for each fraction was then input into GlycoPep DB; and, using the high-resolution MS¹ peaklist of that fraction, all the plausible glycopeptide compositions could be identified. These glycopeptide compositions were then verified manually using MS¹ and MS/MS data.

4.3 RESULTS AND DISCUSSION

CON-S gp140ΔCFI is a potential candidate for HIV/AIDS vaccine, and it is a very heavily N-glycosylated protein with 31 potential glycosylation sites.²⁹ Figure 4-1 shows the CON-S gp140ΔCFI protein sequence with all the potential glycosylation sites highlighted in red. The peptides boxed in green represent all the possible tryptic peptides containing one or more potential glycosylation site(s) produced from this protein, with no missed

cleavages. The glycosylation on this protein has recently been described,²⁸ and in that work, glycosylation analysis was demonstrated to be an effective technique in correlating glycosylation profiles with vaccine efficacy. The work presented here uses this same protein in a case-study detailing the relative merits of offline HPLC, followed by MALDI TOF/TOF MS and online LC-ESI-LTQ-FTICR-MS for glycopeptide analysis. The protein was subjected to typical sample preparation conditions (reduction/alkylation and tryptic digest) and analyzed using two of the most powerful MS techniques; LC/ESI-FTICR-MS and MALDI-TOF/TOF. Figure 4-2 illustrates the analytical protocol employed in this study. After the glycoprotein was digested with trypsin, the total digest was divided into two portions. Each portion was subjected to either capillary LC/ESI-FTICR MS or HPLC fractionation followed by MALDI-TOF/TOF MS analysis. In addition, the reconstituted HPLC fractions collected for MALDI-TOF/TOF analysis were also deglycosylated and reanalyzed by MALDI-TOF/TOF MS. The glycosylation information content (sequence coverage, number and type of glycans) obtained from each MS approach was compared to determine the strengths and weaknesses of the two methods.

4.3.1 Assigning glycopeptide compositions

One of the key challenges in glycopeptide-based MS analysis is assigning compositions to the masses observed in MS¹ data with a high confidence level. This is because it is very possible to assign different

glycopeptide compositions to the same mass, even when mass accuracy is less than 5 ppm.³⁴ Thus, a comparison was undertaken to determine whether LC/ESI-FTICR or MALDI-TOF/TOF MS had advantages in terms of providing the most confirmatory information about the glycopeptide compositions assigned. To perform this comparison, glycopeptide peaks observed from the high-resolution MS¹ data of each instrument were subjected to MS/MS experiments, and the resulting product ions from each technique were used to confirm the glycopeptide compositions assigned based on the high-resolution MS¹ data. Figure 4-3a and b represent MS/MS data from LC/ESI-FTICR and MALDI-TOF/TOF MS respectively of the same glycopeptide observed in both methods. This glycopeptide is used as an example to demonstrate the relative merits of MS/MS analysis from each technique in providing high confidence assignments for the peptide and glycan compositions.

MRVRG IQRNCQHLWRWGTL I LGMLM ICSAAENLWVTVYYGVPVWKEANTT
LFCASDAKAYDTEVHNVWATHACVPTDPNPQE I VLENVTENFNMWKNMV
EQMHEDI I SLWDQSLKPCVKLTPLCVTLNCTNVNVTNTTNNTEEKGE I KN
CSFNITTEIRDKKQKVYALFYRLDVVPIDDNNNSSNYRLINCNTSAITQ
ACPKVSFEP I P IHYCAPAGFA I LKCNDKKFNGTGPCKNVSTVQCTHG I KP
VVSTQLLLNGSLAEEEE I IIRSENI T NNAKTI I VQLNESVE I NCTRPNNT
RKS I RIGPGQAFYATGDI I GD I RQAHCN I SGTKWNKTLQQVAKKLREHFN
NKTI I IFKPSSGGDLE I TTHSFNCRGEFFYCNTSGLFNSTWIGNGTKNNNN
TNDTI I TLPCR I KQ I INMWQGVGQAMYAPP I EGK I TCKSNITGLLLTRDGG
NNNTNETE I FRPGGGDMRDNRSELYKYKVVK I EPLGVAPT KAKLTVQAR
QLLSG I VQQQSNLLRA I EAQQHLLQLTVWGI KQLQARVLAVERYLKDQQL
E I WDNMTWMEWERE I NN YTD I IYSL I EESQNQQEKNEQEL LALDKWASLW
NWFDI TNWLW

Figure 4-1: The protein sequence for CON-S gp140 Δ CFI with all the 31 potential glycosylation sites highlighted in red. The peptides boxed in green represent all the potentially glycosylated tryptic peptides present in this protein with no missed cleavages.

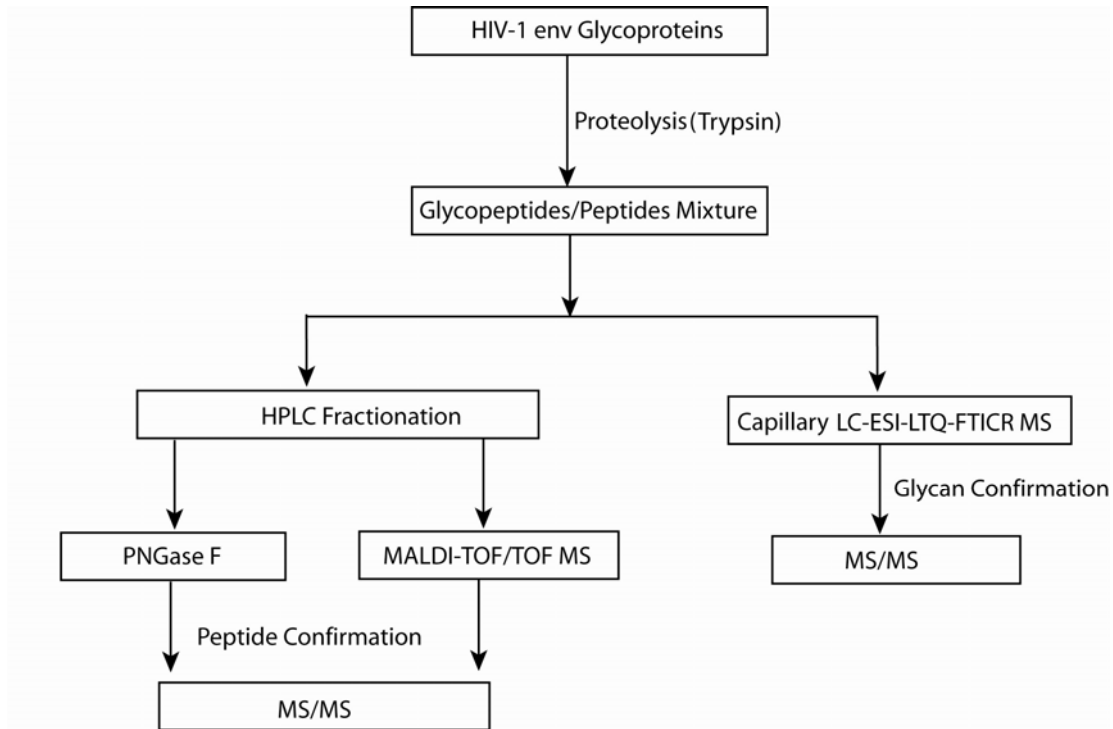


Figure 4-2: Work flow to illustrate the protocol used to analyze CON-S gp140 Δ CFI glycopeptides.

4.3.1.1 MS/MS data from LC/ESI-FTICR MS

Figure 4-3a illustrates an MS/MS spectrum, acquired in the linear ion trap of the LTQ-FTICR mass spectrometer. This spectrum is populated with product ions resulting from glycosidic bond cleavages, which provide information about the glycan moiety attached to the peptide. For instance, the sequential losses of hexose (162 Da), HexNAc (203 Da), and fucose (146 Da) can be identified and used to verify the glycan composition attached to the peptide. In Figure 4-3a, the glycan composition for the glycopeptide at m/z 1477.15 is confirmed by product ions resulting from glycosidic bond cleavage of this glycopeptide. The glycosidic cleavages include sequential losses of nine hexoses (mannose residues), confirming the presence of Man₉, a high mannose type of N-linked glycan. The glycosidic cleavage resulting from a loss of a HexNAc is represented by a square in Figure 4-3a. As indicated in this spectrum, the cleavage of all the glycosidic bonds present in this glycopeptide are observed up to the innermost N-acetylglucosamine residue (HexNAc), which is attached to the peptide moiety of the glycopeptide. This cleavage product corresponds to the Y₁ ion or [Peptide+203+H]⁺. When core fucosylation is present, both [Peptide+203+H]⁺ and [Peptide+349+H]⁺ are observed. The Y₁ ion is a useful characteristic product ion and was observed in all MS/MS data for the

glycopeptides subjected to ESI MS/MS experiments; it provides information about the glycan attachment site. The product ion corresponding to the Y_1 ion can be identified either manually or by simply inputting the MS/MS peaklist for this glycopeptide (m/z 1477.15) into GlycoPep ID, <http://hexose.chem.ku.edu/predictiontable2.php>, which automatically outputs the potential peptide and its corresponding Y_1 ion. (See experimental section) In this case, GlycoPep ID was used to identify the Y_1 ion, which was identified as m/z 1290.74 (singly charged) and m/z 646.08 (doubly charged) and its corresponding peptide moiety, $SN^{453}ITGLLLTR$. Taken together, the glycosidic cleavage product ions explicitly confirm the glycan portion of this glycopeptide, and the Y_1 ion verifies the peptide composition. However, there were no other product ions resulting from cleavage along the peptide backbone; thus further confirmation of the peptide sequence was not feasible.

4.3.1.2 MS/MS data from MALDI-TOF/TOF MS

Figure 4-3b represents MS/MS data obtained from MALDI-TOF/TOF of the same glycopeptide shown in Figure 4-3a. As indicated in Figure 4-3b, fewer fragmentation ions are observed compared to the ones observed in Figure 4-3a. These ions include two sets of cleavage ions at or near the innermost HexNAc residue. The two sets of cleavage ions represent the Y_1 ion ($[Peptide+203+H]^+$) and the $^{0,2}X$ ion ($[Peptide+83+H]^+$). This pair of ions was always observed in all glycopeptides subjected to MALDI-MS/MS

experiments, regardless of the type of N-linked glycans (high-mannose, complex or hybrid) present. Like in MS/MS on the linear ion trap, when core fucosylation is present, the Y_1 ion, corresponding to $[\text{Peptide}+349+\text{H}]^+$, is observed, along with the $^{0,2}\text{X}$ ion. In Figure 4-3b, Y_1 and $^{0,2}\text{X}$ ions are observed at m/z 1290.9 and m/z 1170.8 respectively. Besides the set of cleavage ions, there were no other glycan related cleavage ions observed in the MALDI MS/MS experiments.³³ This is because unlike the low energy CID in the linear ion trap, MALDI-MS/MS is a high energy process that yields predominantly fragmentation ions from peptide bond cleavage.²⁷ As a result, MS/MS of the glycopeptide at m/z 2952.55 yields several y and b ions resulting from peptide bond cleavage. Thus, this technique provides detailed sequence and site-attachment data for the glycosylated peptide but provides minimal information about the glycan moiety.

Overall, MS/MS data acquired from the two high resolution MS techniques, LC/ESI-LTQ-FTICR and MALDI-TOF/TOF provided confirmatory information in that in both methods, the Y_1 ion was always observed. This ion was used as a characteristic ion to identify the peptide moiety of the glycopeptide in question. The remaining mass of the glycopeptide after subtracting the mass of the Y_1 ion can be used indirectly to determine the glycan moiety for that glycopeptide. In MS/MS in the linear ion trap, the characteristic Y_1 ion was always observed as the base peak for glycopeptides containing high-mannose glycan compositions, but was not the

base peak for glycopeptides containing complex or hybrid glycans. In MALDI-MS/MS, in addition to the Y_1 ion, the $^{0,2}X$ ion was also always observed, and either of these two ions formed the base peak. It is worth noting that although these ions identify the peptide moiety of the glycopeptide in question by providing the mass of the peptide, deglycosylation experiments, which identify the peptide with a high degree of confidence since they provide the peptide sequence of the deglycosylated peptide, generally identified the same peptides as identified by the Y_1 ion and the $^{0,2}X$ ions. This increased the confidence level of the identified peptides from both ESI- and MALDI-MS/MS data. Furthermore, for smaller mass ions ($< m/z$ 5000) and strongly ionizing peptides, like arginine-containing tryptic peptides, the peptide sequence could easily be obtained from MALDI MS/MS data without deglycosylation. Therefore, MALDI-TOF/TOF provided a higher confidence level for identifying the peptide moiety than the LC/ESI-FTICR-MS data. However, in terms of the glycan moiety identified by both MS/MS techniques, LC/ESI-FTICR provided a higher confidence level than MALDI-TOF/TOF. When the two MS techniques are used together, extensive information can be obtained both about the peptide sequence and the monosaccharide units contained in the glycan.

4.3.2 Number of glycoforms identified

Table 4-1 shows the number of glycoforms identified at each glycosylation site, detected from both LC/ESI-FTICR and MALDI-TOF/TOF

MS. A complete list of all the assigned glycan compositions can be found in Supplemental Table 4-1, attached. From Table 4-1, it is quite evident that the number of glycans obtained from each glycosylation site differed greatly between the two instruments. For instance, from LC/ESI-FTICR MS data in Table 4-1, we identified 27 different glycan compositions attached to EANTTLFCASDAK peptide whereas, from the same glycosylation site, only four glycan compositions were identified using MALDI-TOF/TOF MS. However, when another glycosylation site is examined, for example, from the peptide LREHFNN³⁶¹K/EHFNN³⁶¹K, 35 different glycan compositions attached to this site were identified using MALDI-TOF/TOF (Table 4-1) whereas from LC/ESI-FTICR MS, only eight different glycan compositions were identified from the same site (Table 4-1). As a result, since the number of glycan compositions identified varied from one glycosylation site to the other between the two instruments, the best glycan population coverage was achieved when the two instruments were used to complement each other.

Figure 4-4 shows a Venn diagram that demonstrates the glycan population coverage for MALDI-TOF/TOF and LC/ESI-FTICR MS. As indicated in this figure, about 130 unique glycan compositions were identified using each of the two MS techniques alone. About 90 identical glycan compositions were identified by both methods. Overall, approximately 350 different glycan compositions were identified from all detected glycosylation sites in CON-S gp140 Δ CFI, using the two high-resolution methods, LC/ESI-

FTICR and MALDI-TOF/TOF MS. These results further support the fact that the best profile for the glycan population present in a complex glycoprotein is best achieved by a combination of both methods.

4.3.3 Identification of the most abundant type of N-linked glycan present

Table 4-1 also shows the most abundant type of N-linked glycans identified from each glycosylation site using both LC/ESI-FTICR MS and MALDI-TOF/TOF MS. (Isomeric structures of the ones shown in Table 4-1 are also possible). All the three types of N-linked glycans, high mannose, hybrid and complex type, were detected from all the identified glycosylation sites; see Supplemental Table 4-1 attached. Although the number of glycans detected at each site using the two methods differed, in most cases, they both provided similar results about the most abundant glycan species present at each site. For example, the total number of glycans found attached to EHFNN³⁶¹K/LREHFNN³⁶¹K using MALDI-TOF/TOF was 35 while from LC/ESI-FTICR MS, only eight were detected. However, regardless of the significant difference in number of glycans detected, the same glycan structure ([Hex₉HexNAc₂]) was identified as the most abundant species in both cases. Additionally, seven of the nine glycosylated tryptic peptides identified by both MS methods produced the exact same glycan composition for the most abundant species. The remaining two glycosylated tryptic

peptides both contained high-mannose glycans, although the exact composition varied slightly between the two instruments. See Table 4-1. This shows that the two instruments provided highly consistent information regarding the most abundant N-linked glycans present at each glycosylation site. From this table, it can also be seen that out of all the glycosylation sites detected by both methods, about 80% of them contained high-mannose N-linked glycans as the most abundant species. As a result, it can be inferred that CON-S gp140 Δ CFI has a high degree of high-mannose N-linked glycan structures.

In summary, the two MS methods used to analyze this sample provided complementary information, both in terms of the number and type of glycosylated peptides detected, and in terms of the glycoforms detected at each site. While the number of glycoforms detected varied, in most cases, each method identified the same type of glycoform as the most abundant species, when the glycopeptide was detectable using both methods.

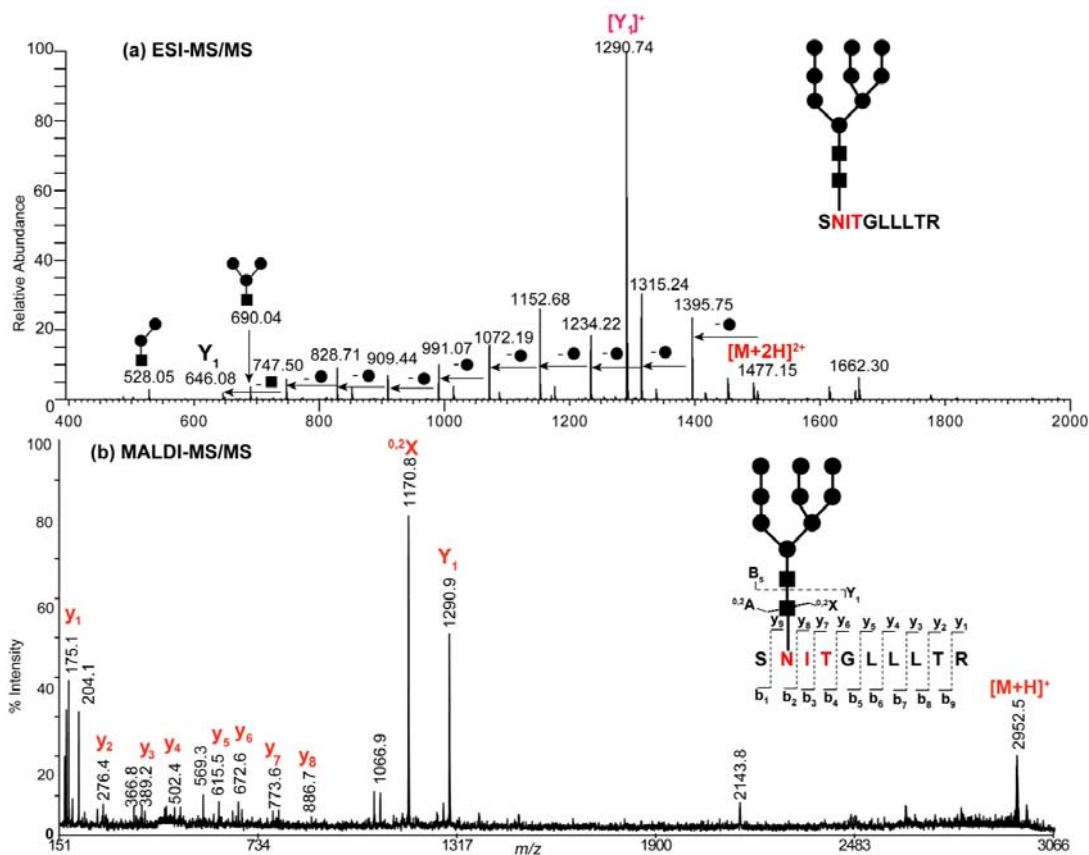


Figure 4-3: A representative example of MS/MS data used to confirm the assigned glycopeptides compositions. (a) Illustrates ESI-MS/MS data for a doubly charged glycopeptide ion at m/z 1477.15. (b) Indicates MALDI-MS/MS data of the singly charged form of the same glycopeptides (m/z 2952.55) as in (a).

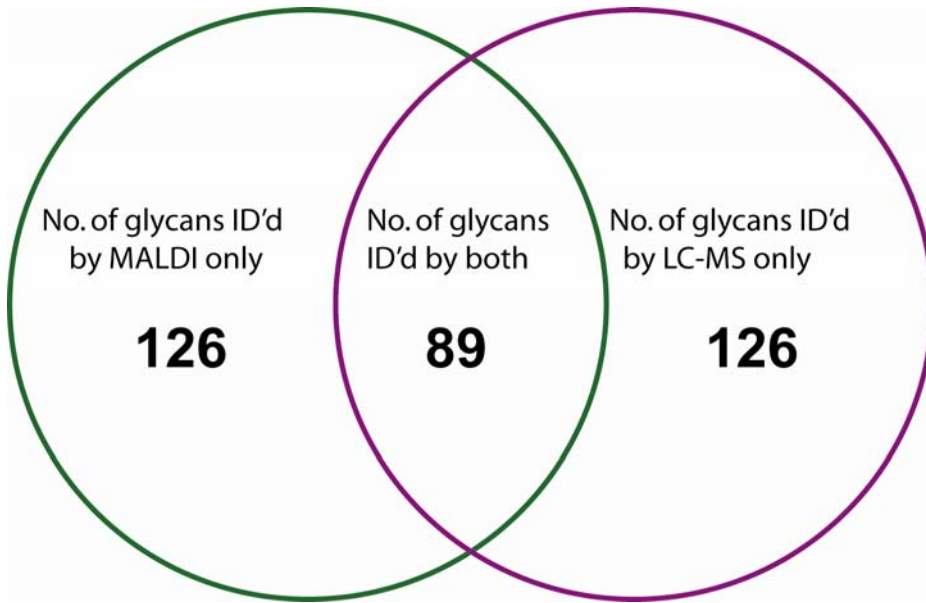
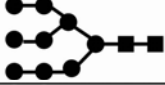
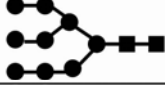
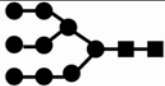
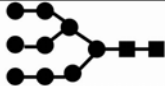
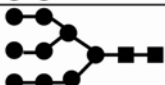

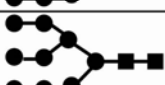


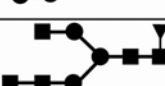


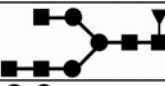

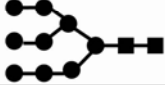

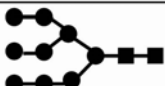
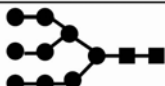
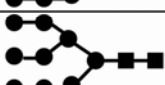
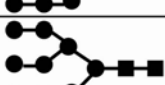
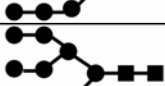
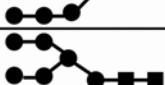

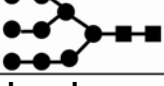


Figure 4-4: Venn diagram indicating the number of glycan population detected by either high resolution LC/ESI-FTICR or MALDI-TOF/TOF or both.

Table 4-1: A summary of type and number of glycoforms identified from CON-S gp140 Δ CFI

Identified Peptide	LC/ESI-FTICR MS		MALDI-TOF/TOF MS	
	No. of glycans	Most abundant structure	No. of glycans	Most abundant structure
FNGTGPCK/ CNDKKFNGTGPCK	4		12	
EHFNNK/ LREHFNNK	8		35	
QAHCNISGTK	5		6	
SENITNNAK	11		1	
NNNNTNDTITLPCR	27		23	
DGGNNNTNETEIFRPG GGDMR	20		45	
LDVVPIDNNNNSSNYR	20		32	
NCSFNITTEIR	32		26	
SNITGLLLTR	21		12	
WNKTLQQVAKK/WNK	2		-	-
EANTTLFCASDAK	27		4	N/A
LINCNTSAITQACPK	24		3	N/A
AYDTEVHNWVATHACV PTDPNPQEIVLENTEN FNMWK	14		-	-
TIIVQLNESVEINCRPN NNTR	-	-	7	
NVSTVQCTHGKIPVVS TQLLLNGSLAEEIIIR	-	-	9	

N/A – Not Applicable, glycopeptide peaks were very low in abundance making it impossible to identify the most abundant type of glycoform present

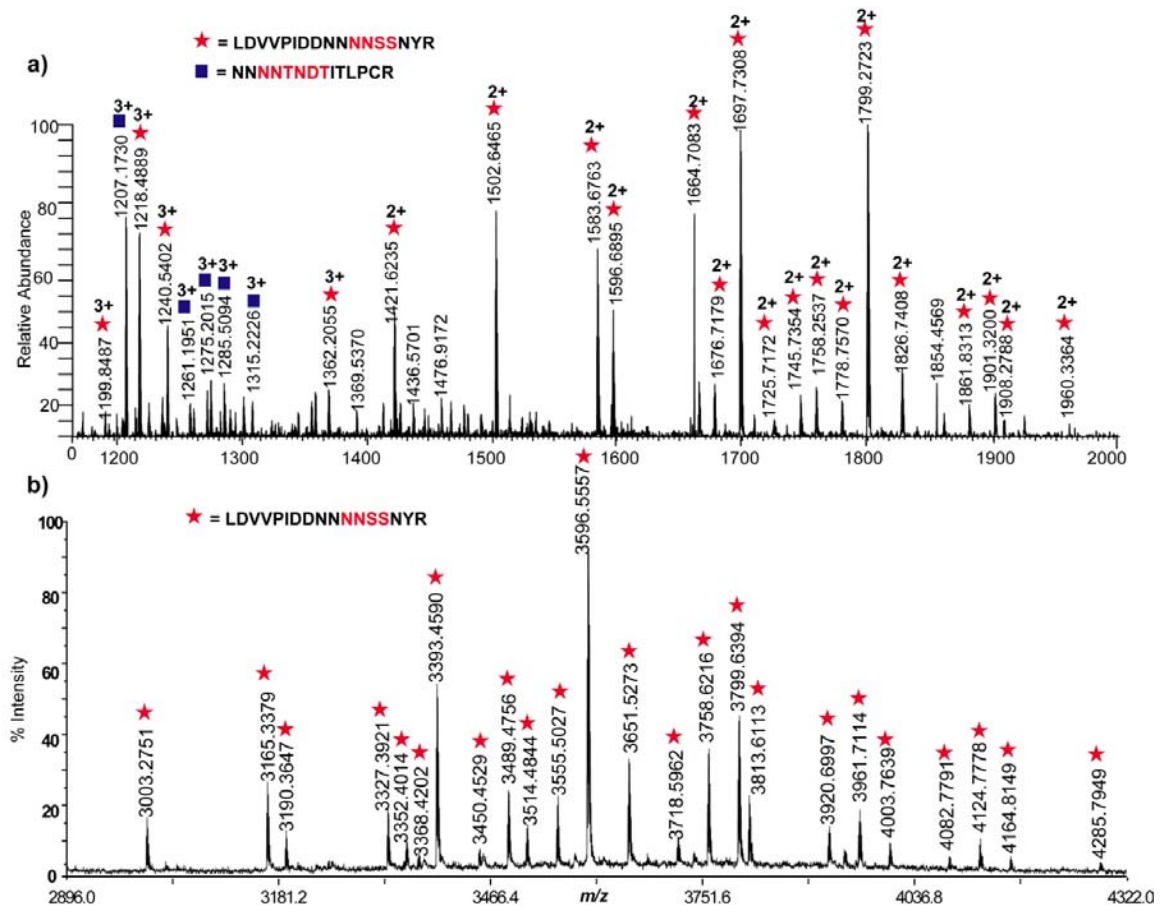


Figure 4-5: Representative examples of MS¹ of high-resolution data containing the same glycopeptide peaks for CON-S gp140ΔCFI acquired on ESI-FTICR and MALDI-TOF/TOF. (a) MS¹ spectrum from ESI (b) MALDI MS¹ spectrum for the similar glycopeptide peaks as ESI MS.

Table 4-2: Glycosylation sites coverage from high-resolution MS

A) Tryptic peptides identified by LC/ESI-FTICR MS and MALDI-TOF/TOF MS	No. of Potential Sites
FN ²³⁷ GTGPCK/CNDKKFN ²³⁷ GTGPCK	1
EHFNN ³⁶¹ K/ LREHFNN ³⁶¹ K	1
QAHCN ³³⁷ ISGTK	1
SEN ²⁸⁰ ITNNAK	1
NNN ⁴¹³ NTN ⁴¹⁶ DTITLPCR	2
DGGNN ⁴⁶⁶ NTN ⁴⁶⁹ ETEIFRPGGGDMR	2
LDVVPIDDNNN ¹⁹⁰ N ¹⁹¹ SSNYR	2
N ¹⁵⁵ CSFN ¹⁵⁹ ITTEIR	2
SN ⁴⁵³ ITGLLLTR	1
EAN ⁴⁸ TTLFCASDAK*	1
LINCN ²⁰¹ TSAITQACPK*	1
B) Unique tryptic peptides detected only in LC/ESI-FTICR MS	
WN ³⁴⁴ KTLQQVAKK/ WN ³⁴⁴ K	1
AYDTEVHNVWATHACVPTDPNPQEIVLEN ⁸⁷ VTENFNMWK	1
EINN ⁶⁴³ YTDIIYSLIEESQNQKEK	1
Unique peptides detected only in MALDI-TOF/TOF MS	
TIIVQLN ²⁹³ ESVEIN ²⁹⁹ CTRPNN ³⁰⁵ NTR	3
N ²⁴⁵ VSTVQCTHGIKPVVSTQLLLN ²⁶⁶ GSLAEEEEIIR	2
DQQLEIWDN ⁶³¹ MTWMEWER	1
C) Undetected tryptic peptides by both LC/ESI-FTICR and MALDI-TOF/TOF MS	
LTPLCVTLN ¹²⁹ CTNVN ¹³⁵ VTN ¹³⁸ TTN ¹⁴¹ NTEEK	4
GEFFYCN ³⁹¹ TSGLFN ³⁹⁷ STWIGN ⁴⁰³ GTK	3

* Peptide sequences detected in low abundance in MALDI-TOF/TOF; verified by deglycosylation with PNGase F and MS/MS on resulting peptides.

4.3.4 Glycosylation sites coverage

Theoretically, digestion of CON-S gp140 Δ CFI with trypsin would produce a total of 19 unique tryptic peptides containing one or more potential glycosylation site(s), (assuming no missed cleavages), which would account for the 31 potential glycosylation sites present in this protein. To determine if any of the two MS techniques could identify all the 31 potential glycosylation sites, (19 tryptic peptides), we examined the number of tryptic peptides and their corresponding number of glycosylation sites detected by each MS technique and then compared the results.

Table 4-2 shows the tryptic peptides (bearing potential glycosylation site(s)) and their corresponding number of glycosylation sites detected from both MS techniques. As shown in this table, from LC/ESI-FTICR MS, a total of 14/19 tryptic peptides each containing one or more potential glycosylation sites were detected, which accounted for 18/31 potential glycosylation sites present in this protein. Figure 4-5a is a representative example of MS¹ data containing glycoforms from two co-eluting tryptic peptides obtained from LC/ESI-FTICR MS. As can be seen from this figure, each tryptic peptide contained various glycoforms. A complete list of glycoforms from each of these tryptic peptides can be found in the Supplemental Table.

In MALDI-TOF/TOF, two analyses were performed in parallel. The first analysis was performed by subjecting each of the reconstituted HPLC fractions to MALDI-TOF/TOF analysis. Figure 4-5b illustrates an example of

MS¹ data from MALDI-TOF/TOF; it contains the same tryptic peptide as that shown in Figure 4-5a, from the LC/ESI-FTICR MS data. The compositions of the glycoforms shown in this figure can be found in Supplemental Table.

In the second MALDI analysis, PNGase F was used to deglycosylate each of the reconstituted HPLC fractions analyzed in the first experiment. This enzyme releases N-linked glycans from the protein, converting the asparagine residues (N) from which the glycans are removed into aspartic acid (D). As a result, a mass shift of 1 Da is expected to occur for every utilized glycosylation site on the peptide, when it is deglycosylated.³⁵ This experiment was used to confirm glycopeptides whose abundance was low in the high-resolution MALDI-TOF/TOF data. For instance, glycosylated tryptic peptides, LINCN²⁰¹TSAITQACPK and EAN⁴⁸TTLFCASDAK, were detected in low abundance in MALDI-TOF/TOF in the first analysis, making it difficult to verify them using MALDI-MS/MS before deglycosylation. However, after deglycosylation (second analysis) these tryptic peptides could be confirmed. More importantly new tryptic peptides were also detected. For example, two tryptic peptides shown in Table 4-2C, LTPLCVTLN¹²⁹CTNVN¹³⁵VTN¹³⁸TTN¹⁴¹NTEEK and GEFFYCN³⁹¹TSGLN³⁹⁷STWIGN⁴⁰³GTK, which contain four and three potential glycosylation sites respectively. These tryptic peptides were also not detected in LC/ESI-FTICR MS, probably because of their high masses, when glycosylated. Another issue with these two peptides that hinders their ionization by MALDI-TOF/TOF is the fact that they are

terminated in lysine. The lysine containing tryptic peptides are known to ionize less efficiently during MALDI analysis than tryptic peptides containing arginine residues³⁶, and their ionization efficiency was even more compromised since they are multiply glycosylated (and thus large and heterogenous). Furthermore, with the high resolution of MALDI-TOF/TOF in the reflectron mode, sensitivity especially for higher masses is lower than for smaller masses, making it more difficult to detect these glycopeptides in the first MALDI-TOF/TOF MS analysis. As a result, it is possible to observe these multiply glycosylated peptides after deglycosylation (second analysis) but not when glycosylated (first analysis). Overall, from high resolution MALDI-TOF/TOF analyses, a total of 14/19 potentially glycosylated tryptic peptides, corresponding to 21/31 potential glycosylation sites, were identified.

In cases where full coverage of all the glycosylation sites is desired, lower resolution MS techniques like MALDI MS analysis in the linear mode can be employed to identify the glycan compositions of the glycosylated peptides uniquely detected in PNGase F experiments; however, this results in glycopeptides mass assignments of lower confidence levels, since the assignments are based on average masses rather than monoisotopic masses and no MS/MS experiments can be performed to confirm the assigned masses.

Based on these results, it is quite evident that using either one of the two high-resolution MS techniques may be inadequate in detecting all the

potential glycosylation sites present in a heavily glycosylated protein like CON-S gp140 Δ CFI. As a result, a comparison was performed to determine what overlap in sequence coverage the two instruments had, and if there was any benefit derived from combining the two MS techniques in terms of the number of glycosylation sites detected. A closer look at these results obtained from both techniques revealed that only 11/19 potentially glycosylated tryptic peptides, accounting for 15/31 potential glycosylation sites, were identified from both methods. (See Table 4-2A) For example, the tryptic glycopeptides shown in Figure 4-3a and b were identified from both MS techniques. The remaining three and six glycosylation sites were uniquely identified from either high resolution LC/ESI-FTICR or MALDI-TOF/TOF MS, respectively and are listed in Table 4-2B. When the numbers of glycosylation sites identified from the two MS techniques are combined, 17/19 tryptic peptides bearing one or more glycosylation sites are identified resulting to a total of 24 of the 31 potential glycosylation sites. This coverage is higher than the 18 or 21 glycosylation sites obtained from either LC/ESI-FTICR or MALDI-TOF/TOF MS, respectively. Approximately 80% glycosylation coverage was obtained when the two high-resolution MS techniques are used together. This implies that it is important to use both techniques in order to increase the probability of detecting as many potential glycosylation sites as possible. Additionally, deglycosylation experiments

followed by low-resolution MALDI-MS methods were necessary to afford 100% coverage in this case.

4.4 Conclusion

This study demonstrates the use of two high-resolution MS techniques; MALDI-TOF/TOF and LC/ESI-FTICR MS, to provide glycosylation information of a potential HIV vaccine candidate, CON-S gp140 Δ CFI, with a high degree of confidence. CID experiments acquired in both instruments indicated that ESI-MS/MS in a linear ion trap provided the most information about the glycan moiety while MS/MS on a MALDI-TOF/TOF provided a higher confidence level for confirming the peptide portion of the same glycopeptide. When used together, the two instruments provided complementary information about the glycopeptide compositions. From the high-resolution data of the two instruments, 14/19 tryptic peptides were obtained from each MS technique accounting for 18/31 and 21/31 potential glycosylation sites in this protein from LC/ESI-FTICR and MALDI-TOF/TOF respectively. When the two instruments were used to complement each other, 24/31 tryptic peptides, accounting for about 80% glycosylation sites coverage, was obtained, indicating that the best glycosylation site coverage is achieved when the two methods are used together.

In terms of glycosylation data, different populations of N-linked glycans comprising of a wide-range of high-mannose, hybrid, and complex types N-linked glycans, were identified and characterized in a glycosylation

site-specific manner. Overall, high-mannose glycans were identified as the most abundant from both MS techniques. Approximately 350 glycopeptide compositions were identified, when data from the two techniques were combined. The information presented in this study provides other researchers with useful insight about what MS methods are most appropriate for glycopeptide analysis, and how those methods can be used synergistically to provide optimal glycosylation coverage and high confidence assignments.

Supplemental Table: Glycopeptide Composition for CON-S gp140 Δ CFI from High-Resolution MS

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI TOF-TOF	
C1	EANTTLFCASDAK	[Hex]3[HexNAc]2	√		
	"	[Hex]3[HexNAc]3	√		
	"	[Hex]3[HexNAc]3[Fuc]1	√		
	"	[Hex]3[HexNAc]4	√		
	"	[Hex]3[HexNAc]4[Fuc]1	√		
	"	[Hex]3[HexNAc]5	√	√	
	"	[Hex]3[HexNAc]5[Fuc]1	√		
	"	[Hex]3[HexNAc]6[Fuc]1	√		
	"	[Hex]4[HexNAc]2	√		
	"	[Hex]4[HexNAc]3	√		
	"	[Hex]4[HexNAc]3[Fuc]1	√		
	"	[Hex]4[HexNAc]4		√	
	"	[Hex]5 [HexNAc]3	√		
	"	[Hex]5[HexNAc]2	√	√	
	"	[Hex]5[HexNAc]3[Fuc]1	√	√	
	"	[Hex]5[HexNAc]3[SO3]1	√		
	"	[Hex]5[HexNAc]4	√		
	"	[Hex]6 [HexNAc]2	√		
	"	[Hex]6[HexNAc]3	√		
	"	[Hex]6[HexNAc]3[Fuc]1	√		
	"	[Hex]6[HexNAc]4	√		
	"	[Hex]7[HexNAc]2	√		
	"	[Hex]8[HexNAc]2	√		
	"	[Hex]9[HexNAc]2	√		
	"	[Hex]4[HexNAc]5	√		
	"	[Hex]5[HexNAc]3[NeuNAc]2	√		
	"	[Hex]6[HexNAc]3[Fuc]1[NeuGc]2	√		
	"	[Hex]7[HexNAc]4	√		
		AYDTEVHNVWATHACVPTDP	[Hex]3[HexNAc]4	√	
		NPQEVVLENVTEHFNMWK			
		"	[Hex]3[HexNAc]5	√	
		"	[Hex]3[HexNAc]6	√	
		"	[Hex]4 [HexNAc]4	√	
	"	[Hex]4[HexNAc]5[Fuc]1	√		
	"	[Hex]4[HexNAc]5[NeuNAc]1	√		
	"	[Hex]5 [HexNAc]3 [Fuc]1	√		
	"	[Hex]5[HexNAc]2	√		
	"	[Hex]5[HexNAc]5	√		

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
C1	AYDTEVHNWVWATHACVPTDP NPQEVVLENVTEHFNMWK	[Hex]5[HexNAc]5[NeuNAc]1	√	
V1-V2	NCSFNITTEIR	[Hex]6 [HexNAc]2	√	
	"	[Hex]7[HexNAc]2	√	
	"	[Hex]8[HexNAc]2	√	
	"	[Hex]9[HexNAc]2	√	
	"	[Hex]3[HexNAc]2	√	√
V1-V2	NCSFNITTEIR	[Hex]3[HexNAc]2[Fuc]1		√
	"	[Hex]3[HexNAc]3		√
	"	[Hex]3[HexNAc]3[Fuc]1	√	√
	"	[Hex]3[HexNAc]4	√	√
	"	[Hex]3[HexNAc]4[Fuc]1	√	√
	"	[Hex]3[HexNAc]5		√
	"	[Hex]3[HexNAc]5[Fuc]1	√	√
	"	[Hex]3[HexNAc]6		√
	"	[Hex]3[HexNAc]6[Fuc]1	√	√
	"	[Hex]4[HexNAc]2	√	√
	"	[Hex]4[HexNAc]3	√	√
	"	[Hex]4[HexNAc]3 [Fuc]1	√	√
	"	[Hex]4[HexNAc]4	√	√
	"	[Hex]4[HexNAc]4 [Fuc]1	√	√
	"	[Hex]4[HexNAc]5[Fuc]1	√	√
	"	[Hex]4[HexNAc]5[Fuc]1[NeuNAc]1	√	
	"	[Hex]5[HexNAc]2	√	√
	"	[Hex]5[HexNAc]3	√	√
	"	[Hex]5[HexNAc]3 [NeuNAc]1	√	
	"	[Hex]5[HexNAc]4 [Fuc]1[NeuNAc]1	√	
	"	[Hex]5[HexNAc]3[Fuc]1	√	√
	"	[Hex]5[HexNAc]4	√	
	"	[Hex]5[HexNAc]4[Fuc]1		√
	"	[Hex]5[HexNAc]4[Fuc]1[NeuNAc]1	√	
	"	[Hex]5[HexNAc]5[Fuc]1		√
	"	[Hex]6[HexNAc]2	√	√
	"	[Hex]6[HexNAc]3	√	√
	"	[Hex]6[HexNAc]3[Fuc]1	√	
	"	[Hex]6[HexNAc]4	√	
	"	[Hex]6[HexNAc]5	√	
	"	[Hex]6[HexNAc]6[Fuc]1	√	√

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
V1-V2	NCSFNITTEIR	[Hex]7[HexNAc]2	√	√
	"	[Hex]7[HexNAc]3	√	
	"	[Hex]7[HexNAc]3Fuc1	√	
	"	[Hex]7[HexNAc]4	√	
	"	[Hex]8 [HexNAc]2	√	√
	"	[Hex]9 [HexNAc]2	√	√
V2	LDVVPIDDNNNNSSNYR	[Hex]3[HexNAc]2		√
	"	[Hex]3[HexNAc]3	√	√
	"	[Hex]3[HexNAc]3[Fuc]1	√	√
	"	[Hex]3[HexNAc]4		√
	"	[Hex]3[HexNAc]4[Fuc]1	√	
	"	[Hex]3[HexNAc]5		√
	"	[Hex]3[HexNAc]5[Fuc]1	√	√
	"	[Hex]3[HexNAc]6[Fuc]1	√	√
	"	[Hex]3[HexNAc]7[Fuc]1		√
	"	[Hex]4[HexNAc]2	√	√
	"	[Hex]4[HexNAc]3	√	√
	"	[Hex]4[HexNAc]3[Fuc]1	√	√
	"	[Hex]4[HexNAc]4[Fuc]1		√
	"	[Hex]4[HexNAc]5[Fuc]1	√	√
	"	[Hex]4[HexNAc]7[Fuc]1		√
	"	[Hex]5 [HexNAc]2	√	√
	"	[Hex]5[HexNAc]3	√	√
	"	[Hex]5[HexNAc]3[Fuc]1	√	√
	"	[Hex]5[HexNAc]4[Fuc]1		√
	"	[Hex]5[HexNAc]5		√
	"	[Hex]5[HexNAc]5[Fuc]1	√	√
	"	[Hex]5[HexNAc]6[Fuc]1		√
	"	[Hex]5[HexNAc]7[Fuc]1		√
	"	[Hex]6 [HexNAc]2	√	√
	"	[Hex]6[HexNAc]3[Fuc]1	√	
	"	[Hex]6[HexNAc]3[NeuNAc]1	√	
	"	[Hex]6[HexNAc]5[Fuc]1		√
	"	[Hex]6[HexNAc]6[Fuc]1		√
	"	[Hex]6[HexNAc]7[Fuc]1		√
	"	[Hex]7[HexNAc]2	√	√
"	[Hex]7[HexNAc]3		√	
"	[Hex]7[HexNAc]6[Fuc]1		√	
"	[Hex]7[HexNAc]7[Fuc]1		√	

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
V2	LDVVPIDNNNNSSNYR	[Hex]8[HexNAc]2	√	√
	"	[Hex]8[HexNAc]4	√	
	"	[Hex]9[HexNAc]2	√	√
V2-C2	LINCNTSAITQACPK	[Hex]3[HexNAc]2	√	√
	"	[Hex]3[HexNAc]3	√	
	"	[Hex]3[HexNAc]3[Fuc]1	√	
	"	[Hex]3[HexNAc]4	√	
	"	[Hex]3[HexNAc]4[Fuc]1	√	
	"	[Hex]3[HexNAc]5[Fuc]1	√	
	"	[Hex]3[HexNAc]6[Fuc]1	√	
	"	[Hex]4[HexNAc]2	√	√
	"	[Hex]4[HexNAc]3	√	
	"	[Hex]4[HexNAc]3[Fuc]1	√	
	"	[Hex]4[HexNAc]5[Fuc]1	√	
	"	[Hex]4[HexNAc]5[Fuc]1[NeuNAc]2	√	
	"	[Hex]5 [HexNAc]2	√	√
	"	[Hex]5[HexNAc]3	√	
	"	[Hex]5[HexNAc]3[Fuc]1	√	
	"	[Hex]5[HexNAc]4[Fuc]1	√	
	"	[Hex]5[HexNAc]4[Fuc]1[NeuNAc]2	√	
	"	[Hex]5[HexNAc]5[Fuc]1	√	
	"	[Hex]6 [HexNAc]2	√	
	"	[Hex]6[HexNAc]3	√	
	"	[Hex]6[HexNAc]3[Fuc]1	√	
	"	[Hex]7[HexNAc]2	√	
	"	[Hex]8[HexNAc]2	√	
"	[Hex]9[HexNAc]2	√		
C2	FNGTGPK	[Hex]3[HexNAc]6[Fuc]2		√
	"	[Hex]4 [HexNAc]4	√	
	"	[Hex]5[HexNAc]4[Fuc]1[NeuNAc]2	√	
	"	[Hex]8[HexNAc]2	√	
	"	[Hex]9[HexNAc]2	√	
	CNDKKFNGTGPK	[Hex]3 [HexNAc]2		√
	"	[Hex]3 [HexNAc]3 [Fuc]1		√
	"	[Hex]3[HexNAc]4[Fuc]1		√
	"	[Hex]3[HexNAc]5[Fuc]1		√
	"	[Hex]4[HexNAc]2		√

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
C2	CNDKKFNGTGPKK	[Hex]4 [HexNAc]3 [Fuc]1		√
	"	[Hex]5 [HexNAc]2		√
	"	[Hex]4[HexNAc]2[Fuc]1		√
	"	[Hex]7[HexNAc]2		√
	"	[Hex]8[HexNAc]2		√
	"	[Hex]9[HexNAc]2		√
	NVSTVQCTHGIKPVVSTQLLL NGSLAEEEEIIR	[Hex]3[HexNAc]2		√
	"	[Hex]3[HexNAc]4[Fuc]1		√
	"	[Hex]3[HexNAc]5[Fuc]1		√
	"	[Hex]4[HexNAc]2		√
	"	[Hex]5[HexNAc]2		√
	"	[Hex]6[HexNAc]2		√
	"	[Hex]7[HexNAc]2		√
	"	[Hex]8[HexNAc]2		√
	"	[Hex]9[HexNAc]2		√
	SENITNNAK	Hex]3[HexNAc]5[Fuc]1	√	
	"	[Hex]4[HexNAc]2		√
	"	[Hex]5 [HexNAc]3	√	
	"	[Hex]5 [HexNAc]3 [NeuNAc]1	√	
	"	[Hex]5[HexNAc]2	√	
"	[Hex]6[HexNAc]2	√		
"	[Hex]6[HexNAc]4	√		
"	[Hex]6[HexNAc]5[Fuc]1	√		
"	[Hex]7[HexNAc]2	√		
"	[Hex]7[HexNAc]4	√		
"	[Hex]8[HexNAc]2	√		
"	[Hex]9[HexNAc]2	√		
C2-V3	TIIVQLNESVEINCTRPNNTR	[Hex]3[HexNAc]2		√
	"	[Hex]3[HexNAc]4[Fuc]1		√
	"	[Hex]3[HexNAc]5[Fuc]1		√
	"	[Hex]4[HexNAc]2		√
	"	[Hex]4[HexNAc]3		√
	"	[Hex]5[HexNAc]2		√
	"	[Hex]7[HexNAc]2		√
V3-C3	QAHCNISGTK	[Hex]4[HexNAc]2		√
	"	[Hex]5[HexNAc]2	√	√

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
V3-C3	QAHCNISGTK	[Hex]6 [HexNAc]2		√
	"	[Hex]6[HexNAc]4[NeuNAc]1	√	
	"	[Hex]7[HexNAc]2		√
	"	[Hex]8[HexNAc]2	√	√
	"	[Hex]9[HexNAc]2		√
	QAHCNISGTKWNK	[Hex]3[HexNAc]8	√	
	"	[Hex]6[HexNAc]3[Fuc]1[NeuGc]2	√	
	C3	WNKTLQQVAKK	[Hex]7[HexNAc]2	√
WNK		[Hex]9[HexNAc]2	√	
EHFNNK		[Hex]3[HexNAc]4[Fuc]1	√	
"		[Hex]3[HexNAc]5[Fuc]1	√	
"		[Hex]4[HexNAc]4[SO3]1	√	
"		[Hex]5[HexNAc]2	√	
"		[Hex]5[HexNAc]3[Fuc]1	√	
"		[Hex]8[HexNAc]2	√	
"		[Hex]9[HexNAc]2	√	
LREHFNNK		[Hex]3[HexNAc]2		√
"		[Hex]3[HexNAc]2[Fuc]1		√
"		[Hex]3[HexNAc]3		√
"		[Hex]3 [HexNAc]3 [Fuc]1		√
"		[Hex]3[HexNAc]4		√
"		[Hex]3[HexNAc]4[Fuc]1		√
"		[Hex]3[HexNAc]5		√
"		[Hex]3[HexNAc]5[Fuc]1		√
"		[Hex]3[HexNAc]6[Fuc]1		√
"		[Hex]4[HexNAc]2		√
"		[Hex]4[HexNAc]2[Fuc]1		√
"		[Hex]4[HexNAc]3		√
"		[Hex]4 [HexNAc]3 [Fuc]1	√	√
"		[Hex]4[HexNAc]4		√
"	[Hex]4[HexNAc]4[Fuc]1		√	
"	[Hex]4[HexNAc]5		√	
"	[Hex]4[HexNAc]5[Fuc]1		√	
"	[Hex]4[HexNAc]6[Fuc]1		√	

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
	LREHFNNK	[Hex]5[HexNAc]2		√
	"	[Hex]5[HexNAc]3		√
	"	[Hex]5[HexNAc]3[Fuc]1		√
	"	[Hex]5[HexNAc]4		√
	"	[Hex]5[HexNAc]4[Fuc]1		√
	"	[Hex]5[HexNAc]5		√
	"	[Hex]5[HexNAc]5[Fuc]1		√
	"	[Hex]6[HexNAc]2		√
	"	[Hex]6[HexNAc]3		√
	"	[Hex]6[HexNAc]3[Fuc]1		√
	"	[Hex]6[HexNAc]5[Fuc]1		√
	"	[Hex]7[HexNAc]2		√
	"	[Hex]7[HexNAc]3		√
	"	[Hex]7[HexNAc]6		√
	"	[Hex]8[HexNAc]2		√
	"	[Hex]8[HexNAc]3		√
	"	[Hex]9[HexNAc]2		√
V4	NNNNTNDITLPCR	[Hex]3[HexNAc]2	√	√
	"	[Hex]3[HexNAc]3[Fuc]1	√	√
	"	[Hex]3[HexNAc]4[Fuc]1	√	√
	"	[Hex]3[HexNAc]4[NeuNAc]1	√	
	"	[Hex]3[HexNAc]5[Fuc]1	√	√
	"	[Hex]4[HexNAc]2	√	√
	"	[Hex]4[HexNAc]2[Fuc]1	√	
	"	[Hex]4[HexNAc]3	√	√
	"	[Hex]4[HexNAc]3[Fuc]1	√	√
	"	[Hex]4[HexNAc]5	√	
	"	[Hex]4[HexNAc]4[Fuc]1		√
	"	[Hex]4[HexNAc]5		√
	"	[Hex]4[HexNAc]5[Fuc]1	√	√
	"	[Hex]4[HexNAc]5[Fuc]1[NeuNAc]2	√	
	"	[Hex]4[HexNAc]6[Fuc]1[NeuNAc]1	√	
	"	[Hex]5 [HexNAc]2	√	√
	"	[Hex]5[HexNAc]3	√	√
	"	[Hex]5[HexNAc]3[Fuc]1	√	√
	"	[Hex]5[HexNAc]4	√	√
	"	[Hex]5[HexNAc]4[Fuc]1	√	
	"	[Hex]5[HexNAc]5[Fuc]1	√	
	"	[Hex]6 [HexNAc]2	√	√

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
V4	NNNTNDITLPCR	[Hex]6[HexNAc]5[Fuc]1	√	
	"	[Hex]6[HexNAc]5[Fuc]2		√
	"	[Hex]6[HexNAc]6[Fuc]1	√	
	"	[Hex]7[HexNAc]2	√	√
	"	[Hex]7[HexNAc]6	√	
	"	[Hex]7[HexNAc]6[Fuc]1	√	
	"	[Hex]8[HexNAc]2	√	√
	"	[Hex]9[HexNAc]2	√	√
	"	[Hex]15 [HexNAc]4		√
	"	[Hex]16 [HexNAc]4		√
	"	[Hex]17 [HexNAc]4		√
	"	[Hex]18 [HexNAc]4		√
C4	SNITGLLLTR	[Hex]3[HexNAc]3[Fuc]1	√	
	"	[Hex]3[HexNAc]4	√	
	"	[Hex]3[HexNAc]4[Fuc]1	√	√
	"	[Hex]3[HexNAc]5	√	√
	"	[Hex]3[HexNAc]5[Fuc]1	√	√
	"	[Hex]3[HexNAc]6[Fuc]1		√
	"	[Hex]4 [HexNAc]3 [Fuc]1	√	
	"	[Hex]4 [HexNAc]2	√	√
	"	[Hex]4[HexNAc]3	√	
	"	[Hex]4[HexNAc]3[Fuc]1	√	
	"	[Hex]4[HexNAc]4	√	
	"	Hex]4 [HexNAc]4 [Fuc]1 [NeuNAc]1	√	
	"	Hex]4 [HexNAc]5		√
	"	[Hex]4[HexNAc]5[Fuc]1[SO3]1	√	
	"	[Hex]5 [HexNAc]2		√
	"	[Hex]5 [HexNAc]3	√	
	"	[Hex]5[HexNAc]3[Fuc]1	√	
	"	[Hex]5[HexNAc]4	√	
	"	[Hex]6 [HexNAc]2	√	√
	"	[Hex]6[HexNAc]3[Fuc]1	√	
	"	[Hex]6[HexNAc]4		√
	"	[Hex]6[HexNAc]5	√	
	"	[Hex]7[HexNAc]2	√	√
"	[Hex]8[HexNAc]2	√	√	
"	[Hex]9[HexNAc]2	√	√	

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
V5	DGGNNNTNETEIFRPGGDM	[Hex]3[HexNAc]2		√
	R			
	"	[Hex]3[HexNAc]2[Fuc]1		√
	"	[Hex]3[HexNAc]3[Fuc]1		√
	"	[Hex]3[HexNAc]4[Fuc]1	√	√
	"	[Hex]3[HexNAc]5[Fuc]1	√	√
	"	[Hex]3[HexNAc]6[Fuc]1	√	√
	"	[Hex]3[HexNAc]7[Fuc]1		√
	"	[Hex]4[HexNAc]2	√	√
	"	[Hex]4[HexNAc]3[Fuc]1	√	√
	"	[Hex]4[HexNAc]4		√
	"	[Hex]4[HexNAc]4[Fuc]1		√
	"	[Hex]4[HexNAc]5	√	√
	"	[Hex]4[HexNAc]5[Fuc]1	√	√
	"	[Hex]4[HexNAc]5[Fuc]1[NeuNAc]1	√	
	"	[Hex]4[HexNAc]6[Fuc]1		√
	"	[Hex]4[HexNAc]6[Fuc]1[NeuNAc]1	√	
	"	[Hex]5[HexNAc]2	√	√
	"	[Hex]5[HexNAc]3[Fuc]1	√	√
	"	[Hex]5[HexNAc]4[Fuc]1		√
	"	[Hex]5[HexNAc]5[Fuc]1	√	√
	"	[Hex]5[HexNAc]5[Fuc]1[NeuNAc]1	√	
	"	[Hex]5[HexNAc]6[Fuc]1		√
	"	[Hex]6 [HexNAc]2	√	√
	"	[Hex]6[HexNAc]3[Fuc]1	√	
	"	[Hex]6[HexNAc]3[NeuNAc]1	√	
	"	[Hex]6[HexNAc]4[NeuNAc]1	√	
	"	[Hex]6[HexNAc]5[Fuc]1	√	√
	"	[Hex]6[HexNAc]5[Fuc]1[NeuNAc]1	√	
	"	[Hex]6[HexNAc]7[Fuc]2		√
	"	[Hex]6[HexNAc]8[Fuc]2		√
	"	[Hex]6[HexNAc]9[Fuc]2		√
	"	[Hex]6[HexNAc]10[Fuc]2		√
	"	[Hex]6[HexNAc]11[Fuc]2		√
	"	[Hex]6[HexNAc]12[Fuc]2		√
	"	[Hex]7[HexNAc]2	√	√
	"	[Hex]7[HexNAc]7[Fuc]2		√
	"	[Hex]7[HexNAc]9[Fuc]2		√
	"	[Hex]7[HexNAc]10[Fuc]2		√
	"	[Hex]7[HexNAc]11[Fuc]2		√

Env Domain	Peptide Sequence	Carbohydrate Composition	LC/ESI-FTICR	MALDI-TOF-TOF
V5	DGGNNNTNETEIFRPGGGDM	[Hex]8[HexNAc]2		√
	R			
	"	[Hex]8[HexNAc]4		√
	"	[Hex]8[HexNAc]6[Fuc]2		√
	"	[Hex]8[HexNAc]7[Fuc]2		√
	"	[Hex]8[HexNAc]8[Fuc]2		√
	"	[Hex]9[HexNAc]2		√
	"	[Hex]9[HexNAc]4		√
	"	[Hex]10[HexNAc]4		√
	"	[Hex]11[HexNAc]4		√
	"	[Hex]12[HexNAc]4		√
	"	[Hex]13[HexNAc]4		√
	"	[Hex]13[HexNAc]5		√
"	[Hex]13[HexNAc]6		√	
TM	EINNYTDIIYSLIEESQNNQEK	Non-glycosylated	√	
	DQQLEIWDNMTWMEWER	Non-glycosylated		√
	DQQLEIWDNMTWMEWER	Non-glycosylated		√

√ - Indicate that the glycoform in question was detected

TM – Transmembrane

Hex – Hexose; HexNAc – N-acetylglucosamine; Fuc – Fucose

4.5 Significance of analyzing CON-Sgp140 Δ CFI

The HIV-1 virus is highly pathogenic, causing one of the deadliest diseases known in human history; therefore, developing a vaccine against this virus is an urgent global priority. So far, considerable efforts have been made towards designing an efficacious HIV vaccine. For a vaccine to be considered as effective, it would have to prevent HIV infection in the vaccinated individuals through eliciting an effective immune response or limiting HIV replication rates, thus delaying/preventing HIV progression to AIDS in the infected individuals. Unfortunately, attempts to design such a vaccine or immunogen have been largely unsuccessful.

One of the major hurdles for developing an effective HIV vaccine is the high level of genetic diversity resulting from its rapid replication and mutational rates. This results in high variability in amino acid sequences of the same HIV-1 group. For instance, HIV-1 main (M) group is the major cause of the HIV pandemic. Within this group there are nine major subtypes or clades, which include A to D, F to H, J, and K. Sequence variability is known to occur between different subtypes and within the same subtype. As a result, it is unrealistic to develop a vaccine based on only one subtype. However, most of the current strategies to develop an HIV vaccine have failed to address the genetic variability of HIV-1 strains. Recently, a new approach was developed that addresses the genetic variability by designing immunogens that are based on “centralized” (ancestral or consensus) HIV

sequences thereby minimizing the genetic gap between different HIV strains. Such an immunogen would be a better representative of contemporary viruses and is expected to elicit neutralizing antibodies against a broader spectrum of viral strains. So far, only a few immunogens have been developed using this approach. One of the most successful immunogens in eliciting neutralizing antibody response from various subtypes of HIV-1 group M is CON-Sgp140 Δ CFI, making this immunogen a potential candidate for HIV vaccine.

Although designing an immunogen with “centralized” sequence is a great advancement towards developing an effective vaccine, it is critically important to analyze the extensive glycosylation pattern on the surface of the immunogen, since it is known to be the key defense mechanism for the virus against immune attack. As a result, to successfully design an efficacious HIV vaccine, one of the initial fundamental steps is to map and profile glycosylation patterns present in HIV envelope proteins and correlate this glycosylation information to their immunological properties. To this end, we have developed mass spectrometric methods to characterize the CON-S gp140 Δ CFI glycosylation pattern. This information will facilitate development of an HIV vaccine that not only optimizes on the peptide sequence but also on the glycan moieties.

4.5.1 CON-S consensus gene design

CON-S, a synthetic group M consensus *env* gene, was constructed by aligning the consensus *env* sequences of group M subtypes A to D, F and G from the 2001 HIV sequence database as described in (http://hiv-web.lanl.gov/content/hiv-db/CONSENSUS/M_GROUP/Consensus.html). The alignment only contained full-length proteins and one sequence from each individual. Besides the hypervariable loops, this consensus has the same regions as a model of the ancestral sequence of the group M that is based on maximum probability phylogenies.³⁷ The hypervariable loop regions (V1, V2, V4, and V5) in the *env* gene evolve by rapid insertion and deletion, whereas the V3 region mainly evolves by point mutation with minimal insertions and deletions. These regions were designed by hand alignments that initially brought potential N-linked glycosylation sites and cysteines into alignment before bringing the repeated sequence motifs within loops into alignment. The V3 hypervariable region was aligned and treated in the same manner as the conserved (C1 to C4) Env regions. Most of the positions of each subtype maintained the same amino acids producing a consensus of consensuses. The resulting consensus contained hypervariable loop sequences of shorter range of lengths than found among natural strains. This was desirable because the shorter hypervariable loops are more likely to expose conserved epitopes that can easily be accessed by neutralizing antibodies.

4.5.2 Expression of recombinant HIV-envelopes

A detailed description of how CON-S was expressed can be found in literature.^{29,37} Briefly, the gene for CON-S *env* was generated by converting its protein sequence into a nucleotide sequence. This was done by utilizing the codon usage of highly expressed human housekeeping genes and de novo synthesized. HIV-1gp140 Envs with the deletion of the cleavage (C) site, fusion (F), and immunodominant (I) region in gp41 hence the name gp140CFI. CON-S gp140 Δ CFI was generated by PCR by introducing a stop codon before the spanning domain (YIKIFIMIVGGLIGLRIVFAVL SIVN). Recombinant vaccinia virus (rVV) expressing CON-Sgp140 Δ CFI were generated and confirmed by PCR and nucleotide analysis after transfection into 293T cells. Recombinant CON-Sgp140 Δ CFI glycoprotein was purified from supernatants of 293T cell cultures using *Galanthus nivalis* lectin-agarose (Vector Labs, Burlingame, CA) column chromatography and stored at -70°C until use.

4.6 References

1. Apweiler, R.; Hermjakob, H.; Sharon, N. *Biochim Biophys Acta*, **1999**, *1473*, 4-8.
2. Moens, S.; Vanderleyden, J. *Arch Microbiol* **1997**, *168*, 169-175.
3. Morelle, W.; Michalski, J.-C. *Curr Pharm Des* **2005**, *11*, 2615-2645.
4. Baenziger, J. U. *Am J Pathol* **1985**, *121*, 382-391.

5. Rudd, P. M.; Wormald, M. R.; Dwek, R. A. *Trends Biotechnol* **2004**, *22*, 524-530.
6. Wang, L.-X. *Curr Opin Drug Discov Devel* **2006**, *9*, 194-206.
7. Geyer, H.; Holschbach, C.; Hunsmann, G.; Schneider, J. *J Biol Chem* **1988**, *263*, 11760-11767.
8. Koch, M.; Pancera, M.; Kwong, P. D.; Kolchinsky, P.; Grundner, C.; Wang, L.; Hendrickson, W. A.; Sodroski, J.; Wyatt, R. *Virology* **2003**, *313*, 387-400.
9. Leonard, C. K.; Spellman, M. W.; Riddle, L.; Harris, R. J.; Thomas, J. N.; Gregory, T. J. *J Biol Chem* **1990**, *265*, 10373-10382.
10. Mizuochi, T.; Matthews, T. J.; Kato, M.; Hamako, J.; Titani, K.; Solomon, J.; Feizi, T. *J Biol Chem* **1990**, *265*, 8519-8524.
11. Back, N. K.; Smit, L.; De Jong, J. J.; Keulen, W.; Schutten, M.; Goudsmit, J.; Tersmette, M. *Virology* **1994**, *199*, 431-438.
12. Cheng-Mayer, C.; Brown, A.; Harouse, J.; Luciw, P. A.; Mayer, A. J. *J Virol* **1999**, *73*, 5294-5300.
13. Kwong, P. D.; Doyle, M. L.; Casper, D. J.; Cicala, C.; Leavitt, S. A.; Majeed, S.; Steenbeke, T. D.; Venturi, M.; Chaiken, I.; Fung, M.; Katinger, H.; Parren, P. W. I. H.; Robinson, J.; Van Ryk, D.; Wang, L.; Burton, D. R.; Freire, E.; Wyatt, R.; Sodroski, J.; Hendrickson, W. A.; Arthos, J. *Nature* **2002**, *420*, 678-682.
14. Reitter, J. N.; Means, R. E.; Desrosiers, R. C. *Nat Med* **1998**, *4*, 679-684.
15. Wyatt, R.; Kwong, P. D.; Desjardins, E.; Sweet, R. G.; Robinson, J.; Hendrickson, W. A.; Sodroski, J. G. *Nature* **1998**, *393*, 705-711.
16. Zhu, X.; Borchers, C.; Bienstock, R. J.; Tomer, K. B. *Biochemistry* **2000**, *39*, 11194-11204.
17. Yeh, J.-C.; Seals, J. R.; Murphy, C. I.; van Halbeek, H.; Cummings, R. D. *Biochemistry* **1993**, *32*, 11087-11099.
18. Morelle, W.; Canis, K.; Chirat, F.; Faid, V.; Michalski, J.-C. *Proteomics* **2006**, *6*, 3993-4015.

19. Wada, Y.; Azadi, P.; Costello, C. E.; Dell, A.; Dwek, R. A.; Geyer, H.; Geyer, R.; Kakehi, K.; Karlsson, N. G.; Kato, K.; Kawasaki, N.; Khoo, K.-H.; Kim, S.; Kondo, A.; Lattova, E.; Mechref, Y.; Miyoshi, E.; Nakamura, K.; Narimatsu, H.; Novotny, M. V.; Packer, N. H.; Perreault, H.; Peter-Katalinic, J.; Pohlentz, G.; Reinhold, V. N.; Rudd, P. M.; Suzuki, A.; Taniguchi, N. *Glycobiology* **2007**, *17*, 411-422.
20. Budnik, B. A.; Lee, R. S.; Steen, J. A. J. *Biochim Biophys Acta* **2006**, *1764*, 1870-1880.
21. Mechref, Y.; Novotny, M. V. *Chem Rev* **2002**, *102*, 321-369.
22. Wuhrer, M.; Koeleman, C. A. M.; Hokke, C. H.; Deelder, A. M. *Anal Chem* **2005**, *77*, 886-894.
23. Carr, S. A.; Huddleston, M. J.; Bean, M. F. *Protein Sci* **1993**, *2*, 183-196.
24. Stephens, E.; Maslen, S. L.; Green, L. G.; Williams, D. H. *Anal Chem* **2004**, *76*, 2343-2354.
25. Burlingame, A. L. *Curr Opin Biotechnol* **1996**, *7*, 4-10.
26. Kuroguchi, M.; Nishimura, S.-I. *Anal Chem* **2004**, *76*, 6097-6101.
27. Wuhrer, M.; Catalina, M. I.; Deelder, A. M.; Hokke, C. H. *J Chromatogr B Analyt Technol Biomed Life Sci* **2007**, *849*, 115-128.
28. Eden, P. G.; Irungu, J.; Zhang, Y.; Dalpathado, D.S; Liao, H-X; Sutherland, L.L;Alam, S.M; Haynes, B. F; Heather, D. *Submitted to Virology* 2007.
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. *Virology* **2006**, *353*, 268-282.
30. Cutalo, J. M.; Deterding, L. J.; Tomer, K. B. *J Am Soc Mass Spectrom* **2004**, *15*, 1545-1555.
31. Go, E. P.; Rebecchi, K. R.; Dalpathado, D. S.; Bandu, M. L.; Zhang, Y.; Desaire, H. *Anal Chem* **2007**, *79*, 1708-1713.
32. Irungu, J.; Go, E. P.; Dalpathado, D. S.; Desaire, H. *Anal Chem* **2007**, *79*, 3065- 3074.

33. Wuhrer, M.; Hokke, C. H.; Deelder, A. M. *Rapid Comm Mass Spectrom* **2004**, *18*, 1741-1748.

34. Irungu, J.; Dalpathado, D. S.; Go, E. P.; Jiang, H.; Ha, H.-V.; Bousfield, G.R.; Desaire, H. *Anal Chem* **2006**, *78*, 1181-1190.

35. Carr, S. A.; Roberts, G. D. *Anal Biochem* **1986**, *157*, 396-406.

36. Brancia, F. L.; Oliver, S. G.; Gaskell, S. J. *Rapid Comm Mass Spectrom* **2000**, *14*, 2070-2073.

37. Gao, F.; Weaver, E. A.; Lu, Z.; Li, Y.; Liao, H.; Ma, B.; A., S. Munir; Scearce, R. M.; Sutherland, L. L.; Yu, J.; Decker, J. M.; Shaw, G. M.; Montefiori, D. C.; Korber, B. T.; Hahn, B. H.; Haynes, B. F. *J. Virol* **2005**, *79*, 1154-1163.

CHAPTER 5

Conclusion and Future Directions

The work described herein focused on developing mass spectrometric methods to characterize glycans in different glycoproteins in a glycosylation site-specific fashion. This approach is highly efficient and allows structural elucidation of both the glycans and their attachment site in a single-MS experiment. A complete characterization of glycans from two important classes of glycoproteins, pituitary glycoproteins and HIV envelope glycoproteins, was performed in a glycosylation site-specific fashion.

The glycans present in pituitary glycoproteins are known to contain unusually high content of terminal residues such as sulfate groups and sialic acid. However, the precise degree of sulfation or sialylation in different glycosylation sites in these glycoproteins and the roles that these residues play are still not well understood. This is mainly due to the acidity and lability of these groups creating a big analytical challenge that has greatly influenced the analysis of these species. Consequently, developing efficient and sensitive analytical techniques that are capable of identifying and characterizing negatively charged glycans in a glycosylation site-specific manner are highly desirable, in order to facilitate the understanding of their biological significance.

Different mass spectrometric methodologies for characterizing glycans containing these terminal residues in a glycosylation site-specific manner were developed. These methods were successfully applied in characterizing all the three pituitary hormones (LH, FSH and TSH). The results provided herein are specifically from the analysis of eFSH and eTSH. To characterize the glycan structures on these glycoproteins, a non-specific enzyme was utilized to generate small glycopeptides that are easier to separate. However, analysis of these glycopeptides can be challenging since it involves simultaneous analysis of two unknowns; the peptide and the glycan portions.

To facilitate identification of the peptide portion, a web-based tool was developed. This tool, known as GlycoPep ID, identifies the peptide portion of negatively charged glycopeptides generated from a non-specific enzyme (proteinase K) by predicting the characteristic product ion observed in (-)MS/MS data of these glycopeptides that corresponds to the peptide portion. The versatility of this method was demonstrated by identifying the peptide moieties of glycopeptides from two different glycoprotein hormones, FSH and TSH, that were exclusively sialylated or sulfated, or were both sialylated and sulfated. A total of 27 peptide moieties were correctly identified by GlycoPep ID and validated using data from a combination of Edman chemistry and high resolution FTICR-MS analysis. This technique represents an important advance in glycosylation profiling because it solves one of the most difficult

problems of using a non-specific enzyme in glycopeptide analysis:

Determining where the enzyme cleaved the protein.

While (-) MS/MS data was useful in identifying the peptide moiety, it provided very minimal glycan structural information, and the amount of information in the spectra vary, depending on the number of SO₃ groups and the charge state of the ion. To overcome this challenge, an ion-pairing approach was developed. This approach utilizes a basic tripeptide to non-covalently interacting with the sulfate group of the sulfated glycopeptide thereby stabilizing it, promoting dissociation pathways that provide more informative product ions. The resulting ion-pair complexes are analyzed using (+) MS/MS to provide structural information on the glycan portion of these glycopeptides. All the sulfated glycopeptides from eTSH were characterized using this approach. The results clearly demonstrated the efficacy of using ion-pairing MS/MS to fully characterize sulfated glycopeptides in a glycosylation site-specific fashion, an approach that is complementary and in most cases superior to (-)MS/MS analysis. The ion-pairing approach provided a wealth of structural information about the glycan portion in addition to being useful for identifying the peptide moiety. The information obtained from MS/MS of the ion-pair complexes was independent of the number of SO₃ groups present or the charge state of the ion and can be used to determine the branching, sequence, and type of N-glycan present in a sulfated glycopeptide.

Although mass spectrometry is widely used for glycoprotein analysis, so far there is no consensus as to which mass spectrometric approach is most suitable or would provide the most glycosylation information with a high degree of confidence. As a result, an investigation was conducted using two high-resolution MS techniques; MALDI-TOF/TOF and LC/ESI-FTICR MS, to provide glycosylation information of a potential vaccine candidate for the HIV virus, CON-S gp140 Δ CFI. CID experiments acquired in both instruments indicated that ESI-MS/MS in a linear ion trap provided the best confidence level for confirming the glycan moiety while MS/MS on a MALDI-TOF/TOF provided a higher confidence level for confirming the peptide portion of the same glycopeptide. When used together, the two instruments provided glycopeptide composition assignments of very high confidence level. In terms of glycosylation data, different populations of N-linked glycans comprising of a wide-range of high-mannose, hybrid, and complex types N-linked glycans, were identified and characterized in a glycosylation site-specific manner. Overall, the high-mannose glycans were identified as the most abundant glycoforms from both MS techniques. Approximately 350 glycopeptide compositions were identified, when data from the two techniques were combined. The information presented in this study provides other researchers with useful insights about what MS methods are most appropriate for glycopeptide analysis, and how those methods can be used

synergistically to provide optimal glycosylation coverage and high confidence assignments.

Future Directions

Unlike the proteomics and glycomics fields where methods of analysis are well established, in the glycoproteomics field, methods of analysis are still under development. The work presented herein makes a significant contribution in advancing this field by solving several problems that have been major hurdles in glycoprotein analysis. These methods can be applied in future studies in fully characterizing HIV envelope proteins. Although a characterization of glycans in an HIV envelope protein was described herein, the mass spectral data analyzed from this protein mainly focused on the positive ion mode data. However, analysis in the positive ion mode mainly favors neutral species while the signal for negatively charged species is suppressed by the strongly ionizing neutral species. This is evident from the data reported, which indicated that less than 10% of all the glycans present in this HIV envelope protein are negatively charged. Moreover, previous studies have suggested that approximately 40% of the glycans found on HIV env proteins contain negatively charged residues. This implies that the 350 glycan compositions characterized in our studies only represent 60% of the total glycans present while the rest remain to be characterized. Since the presence of negatively charged species can act as points of interactions or as specific recognition markers for receptor binding, characterizing glycans

capped with these residues in HIV envelope proteins is of great importance. The ion pairing approach could be used to stabilize the sulfate group, when present, to facilitate structural analysis of these glycans; whereas GlycoPep ID can be used to simplify mass spectral data analysis.

In addition, since the envelope proteins in the HIV virus are known to evolve rapidly during infection and HIV disease progression by changing the glycan position, number, and structures, resulting in new virus strains that successfully escape any immune attack, a thorough investigation of glycosylation on different HIV envelope proteins originating from the same or different strains is required. Thus future work will include such a study that would focus on identifying the conserved glycosylation sites in HIV proteins which can eventually be targeted during HIV vaccine development. Since the glycosylation process is host cell dependent, a careful evaluation should be performed to determine the relative consistency of glycosylation patterns found in different mammalian cells that are typically used to propagate HIV viruses. Once the overall glycosylation patterns are defined, the mammalian cell line containing the most conserved glycosylation pattern would be selected for vaccine development purposes. Further studies will be performed to map glycosylation patterns of different HIV envelope proteins expressed in the selected cell line to identify the conserved glycosylation sites or glycosylation sites that are unutilized or contain small glycans that are easily accessible to neutralizing antibodies. The glycosylation information

obtained therewith, when correlated with the immunological response of these proteins, will greatly facilitate development of an effective HIV vaccine.

In addition, future studies would also focus on facilitating development of new types of HIV vaccines. Several studies have reported striking disparities between the glycans of HIV-infected and healthy cells. For example, the presence of dense high mannose glycans structures on gp120 is a unique feature of infected cells that is not typical of healthy cells. The feasibility of exploiting such unique features to develop an HIV vaccine was demonstrated by the discovery of 2G12, one of the broadly neutralizing human antibodies to HIV. Another way to exploit these distinct glycan features would be to develop a glycopeptide-based HIV vaccine. Since glycans are known to be poor immunogens, the invention of this type of vaccine will be able to take advantage of the unique features of glycans and also target the conserved peptide moiety where the glycans are attached. This will not only lead to a vaccine that has better immunogenicity than the glycan-based vaccine, but may also unravel new epitopes that could potentially be used as targets for neutralizing antibodies. However, the progress of these studies will not only require a fundamental knowledge of the glycan structures on HIV envelope glycoproteins but also their specific locations on the protein. Consequently, the developed mass spectrometric methodologies presented herein could be used facilitate such studies.