MS/MS ANALYSIS OF IgG3 DISULFIDE BONDS AND DEVELOPMENT OF A NOVEL TOOL TO ASSESS ALGORITHMS THAT ASSIGN GLYCOPEPTIDE CID DATA

By

Jude C. Lakbub

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

_____
Chairperson: Dr. Heather Desaire

_____
Dr. Minae Mure

_____
Dr. Michael Johnson

_____
Dr. Michael Wang

_____
Dr. Robert Dunn

Date Defended: September 1st, 2017

The dissertation committee for Jude C. Lakbub
certifies that this is the approved version of the following dissertation:

_MS/MS_ ANALYSIS OF IgG3 DISULFIDE BONDS AND DEVELOPMENT OF A NOVEL TOOL TO ASSESS ALGORITHMS THAT ASSIGN GLYCOPEPTIDE CID DATA

Chairperson: Dr. Heather Desaire

Date approved: September 1st, 2017

Abstract

With the rapidly increasing use of proteins as biotherapeutics to treat diseases, the characterization of these large molecules using mass spectrometry has become a highly attractive field of research. A particular area of research is the identification and characterization of protein post-translational modifications. Disulfide bonds and glycosylation are among the most critical protein post-translational modifications (PTMs), as they play vital roles in maintaining the proper protein folding, structure, and functions. These two PTMs are particularly important in the development and characterization of monoclonal antibody-based drugs, which are the most prevalent protein therapeutics in the market. Among the four classes of immunoglobulins (IgG's), the disulfide connectivity of IgG1, IgG2 and IgG4 have been effectively studied, and IgG2 and IgG4 have been shown to have disulfide bond-mediated isomers due to alternative disulfide bond connectivity. However, no studies to investigate the presence of disulfide related isomers in IgG3 have been done. In this dissertation, high resolution mass spectrometry is used map the disulfide bond connectivity in IgG3 in order to investigate the presence of disulfide-mediated isomers. The results indicate that no such isomers exist for endogenous IgG3 antibodies.

The development of a novel glycoproteomics software, *Glycopep Decoy Generator* (Tool 1), and the generation of a large dataset of manually assigned CID spectra (Tool 2) from diverse glycopeptide compositions also are described herein. The decoy generator generates abundant decoys for any target glycopeptide composition, and when it is used along with the dataset of CID spectra, the accuracy of glycopeptide scoring algorithms can be readily determined. The tools were used to assess GlycoPep Grader, a scoring algorithm that assigns glycopeptides to

CID spectra. The results indicate that GlycoPep Grader has some weaknesses in scoring spectra from fucosylated glycopeptide compositions. These weaknesses could not be easily identified without the aforementioned tools. In order to address GlycoPep Grader's limitations, a thorough investigation of the root cause of its weaknesses is carried out, and potential updates that could improve the software are proposed.

Acknowledgements

I could not have made it this far in my education without the Grace of God Almighty. I thank God for my life and the success I have had at all levels of my education. I trust Him to see me through my career as a scientist. and to help me be a better scientist.

Thank you to the Chemistry faculty of the University of Kansas for teaching me the science I needed to effectively carry out my research projects. My utmost thanks go to my supervisor, Dr. Heather Desaire, for being an awesome mentor. You always go above and beyond for your students; you took time out of your busy schedule to teach us writing skills and the basics of mass spectrometry even though we had taken mass spec class. Thank you! I am aslo very grateful that you let me spend one year doing research in a pharmaceutical industry, where I put to practice what I learned in your group and I also expanded into the field of released glycan analysis. As I begin my career, I can only hope to emulate your integrity and dedication to science, and to be a fruitful branch of the Desaire Research Tree.

I would also like to thank the current and past members of the Desaire Research Group who made life fun in the group. My sincere appreciation to Dr. Daniel Clark, Dr. Zhikai Zhu, and Dr. Eden Go for training me when I joined the group. Xiaomeng Su and David Hua, thank you for your collaboration in developing the software I needed to complete some of my projects. Thank you to Kasun, Josh, Milani, and Wenting for the time we spent together. I hope you learned from me as much as I did from you.

Finally, I want to thank my family for their love and support. I am grateful for my wife, Adele Botan, who is a great partner and friend. Thank you for your hard work in keeping our

family going. You are a blessing to me, and I thank God that I have you as a life partner. To my parents, John and Frida Lakbub, I am very grateful for your fervent prayers and financial support throughout the various stages of my education. I still have some of the *Express Union* receipts for money you sent to me between 2004 and 2008 while I was at the University of Buea. Thank you very much! Lastly, I want to thank my siblings, Shirlie, Gideon, Ernestine, Judith, Romeo, Elmang, Linus, Denis, Paula, Kintim, Kwasinwi, and Mercy. I really missed you all during these years I spent thousands of miles away from you. You are always in my mind and heart, and you have been the secret source of strength that galvanized me to work harder in order to make you proud. I love you all very dearly, and I wish the best for each of you. I will always be there for you when you need me.

Table of Contents

Chapter 1: Introduction

1.1 Protein Disulfide Bonds

1.1.1 Formation and Importance

Disulfide bonds are post-translational modifications in proteins formed between the sulfur atoms of two cysteine (Cys) residues during the biosynthesis of the proteins in the cell. The formation of the covalent bond results from oxidation of the free thiol (-SH) side chains of the Cys residues, primarily catalyzed by enzymes, including protein disulfide isomerase (PDI) and endoplasmic reticulum oxidoreductin 1protein (Ero1p).[1,2] A significant number of proteins contain disulfide bonds. Based on the known tertiary structures of 817 plasma proteins that contain 4594 disulfide bonds, Butera *et al* [3] approximated the ratio of protein-to-disulfide bond at 1:5. Hence, the approximately 2000 plasma proteins identified by Farrah *et al* [4] would contain about 10,000 disulfide bonds, representing an enormous number of disulfide bonds in plasma proteins, alone. Disulfide bonds are important in protein folding, and they have both structural and functional roles in the proteins.

Structurally, disulfide bonds ensure proper folding of proteins, can lead to structural isoforms,[5] and they stabilize the native high-order conformations of the proteins that are necessary to execute their biological functions.[6,7] The concept of disulfide engineering is, therefore, an attractive choice in biotechnology as non-native disulfide bonds can be engineered into proteins to increase the protein stability. For instance, some proteins that initially lacked disulfide bonds have been shown to be more stable with engineered disulfides,[8,9] and the stability of proteins with native disulfide bonds increased with the introduction of additional disulfide

bonds.[10-13] In some cases, engineered disulfide bonds increased the protein's half-life,[8,14] reduced self-aggregation,[15] and decreased immunogenicity.[16] Besides proteins, peptides with engineered disulfide bonds have also shown increased stability and half-life.[17,18] Although the above examples demonstrate the benefits of additional disulfide bonds in proteins and peptides, not all engineered disulfide bonds produce the expected increase in protein stability.[19]

Some disulfide bonds, known as allosteric disulfides, are responsible for effective biological functions of proteins, and the cleavage of such bonds would lead to a change in the protein activity.[20] The functional roles of allosteric disulfide bonds in blood and cancer cells have been extensively reviewed by Hogg *et al*.[3,21] A recent report showed that reduction and alkylation of some disulfide bonds in rituximab and trastuzumab, IgG1-based drugs, increased the binding affinity of the modified drug to some Fc gamma receptor isotypes,[22,23] but also led to decreased binding to other Fc gamma receptors.[23] In some cases, mutation of Cys residues involved in disulfide bonds may have no effect on the protein's biological activity, as is the case of an IgG2 antibody where Cys to Ser mutations led to structural changes but had no impact on the binding of the protein to receptors and to complement C1.[24]

Mapping the disulfide connectivity pattern in proteins, therefore, provides important information for research pertaining to protein stability, structure-function relationships, and any disulfide-mediated isoforms of proteins. In addition, disulfide bond characterization is of high importance during the development of biopharmaceuticals to ensure the safety and potency of biologics, which have increased dramatically in the drug market in recent years. Hence, there is an increasing demand for efficient analytical methods for accurate characterization of disulfide bonds in proteins, particularly therapeutic proteins.

1.1.2 Disulfide Bonds in Biotherapeutics

Protein disulfide bond characterization has become even more important in biopharmaceutical industries, due to the increasing use of recombinant proteins as biotherapeutics (biologics) for the treatment of diseases such as cancer, arthritis, asthma, and diabetes;[25,26] and as vaccines against various diseases.[27,28] These biologics are from a vast array of protein classes, including hormones,[25,29] monoclonal antibodies,[25,26,30] and growth factors.[31,32] All these classes of biomolecules contain proteins that are disulfide bonded. Among these proteins, immunoglobulin gamma (IgG) antibodies are highly disulfide bonded (between 16 to 25 disulfide bonds, depending on the type of IgG),[33] and IgG-based therapeutics are the most prevalent in the market. As such, disulfide bonds are one of the many critical quality attributes (CAQs) of antibody-based drugs that have to be monitored throughout their development stages, as disulfide bond reduction and scrambling can occur in biotherapeutics during manufacturing[34,35] and storage.[36,37] The presence of free Cys (reduced disulfide bonds) have been reported during manufacturing, and they can lead to formation of non-native disulfide bonds and aggregates,[38,39] and possibly cause immunogenicity and loss of biological activity of the biotherapeutics. Hence, disulfide bond characterization is necessary during biologic development to confirm the correct disulfide connectivity and to verify the presence, or absence, of disulfide bond variants in order to ensure the safety and efficacy of the drugs. In addition, regulatory agencies require comprehensive characterization of the disulfide bond pattern in biomolecules, in order to meet the quality-by-design (QbD) requirements for biologics,[40,41] hence the need for effective methods to map disulfide bonds in biotherapeutics.

1.2 Analytical Methods for Disulfide Bond Characterization

Several methods for disulfide bond analysis in proteins have been developed using a variety of analytical techniques, including NMR,[42-44] X-ray crystallography,[45] Edman degradation,[5,46] diagonal paper electrophoresis,[47,48] and liquid chromatography coupled to mass spectrometry (LC-MS),[49-51] which is the focus of this review. NMR and X-ray crystallography provide information about disulfide bonds at the molecular level, but they require large amounts of highly pure samples, and they are not typically used for disulfide bond mapping. Traditional methods, such as Edman degradation and diagonal paper electrophoresis, were the prominent methods for disulfide bond mapping in the early 1960s, although Edman degradation methods (in combination with mass spectrometry) were still sparingly used in the late 2000s.[5,46] With the advent of mass spectrometry, LC-MS and tandem mass spectrometry (LC-MS/MS) methods have become the go-to methods for disulfide bond mapping in proteins.

Bottom-up mass spectrometry is the most widely used method for disulfide bond analysis in proteins. The attractive aspects of the bottom-up approach include the availability of a variety of enzymes to digest the large biomolecules into small pieces (peptides containing intact disulfide bonds) that are easier to analyze, several soft ionization techniques, complementary fragmentation techniques, and the ability to couple mass spectrometers with LC systems for separation of the enzymatic digests prior to MS analysis. Despite the wide use of mass spectrometry for disulfide bond analysis, disulfide bond characterization is still a challenging task, especially for proteins with high Cys content and complex disulfide connectivity, and for determination of low levels of disulfide bond scrambling. Hence, numerous methods for disulfide bond analysis have been developed, in part, to address these challenges.

In the following sections, we review the recent bottom-up mass spectrometric methods for disulfide bond analysis and provide important considerations for the steps involved. A number of reviews for disulfide bond analysis containing additional methods that were developed prior to 2007 have been reported.[49-51] Herein, we begin by taking an in-depth look at sample preparation, which is a key step in the successful mapping of native and alternative disulfide bonds in proteins. We provide several important tips to prevent the introduction of disulfide artifacts (or scrambling) during sample preparation. Finally, we discuss the fragmentation characteristics of disulfide-linked peptides upon subjection to various mass spectrometric dissociation techniques that are important for disulfide bond mapping and describe recent MS-based disulfide bond characterization methods that have been developed within the past decade. Researchers involved in method development for protein characterization can use the information herein to facilitate development of new MS-based methods for protein disulfide bond analysis. In addition, individuals doing biotherapeutics characterization, especially disulfide bond mapping in antibodies, can use this review article to choose best strategies for disulfide bond assignment of their biologic products.

1.3 Sample Preparation for Bottom-up Mass Spectrometric Disulfide Bond Analysis

1.3.1 Overview of Sample Preparation Methods

Sample preparation is a critical step in bottom-up mass spectrometric disulfide bond analysis. In general, the analysis is usually done using one of two ways: non-reduced (intact) analysis or reduced/intact analysis. The sample preparation workflow for the non-reduced approach, which is the most commonly used method, is shown in Figure 1. This approach requires proteolytic digestion of the protein without disulfide bond reduction, and the disulfide

linked peptides are investigated to decipher the disulfide connectivity in the protein.[52-54] For the intact/reduced approach, two batches of enzymatically digested samples are prepared, one with the disulfide bonds intact (same as the previous approach) and the other with reduced disulfide bonds. The disulfide connectivity of the protein is determined by comparing the peptide map profiles (LC profiles) of the two sample batches.[55,56] The disulfide bond analysis methods which are based on these sample preparation approaches are discussed fully in Section 1.4 of this chapter.

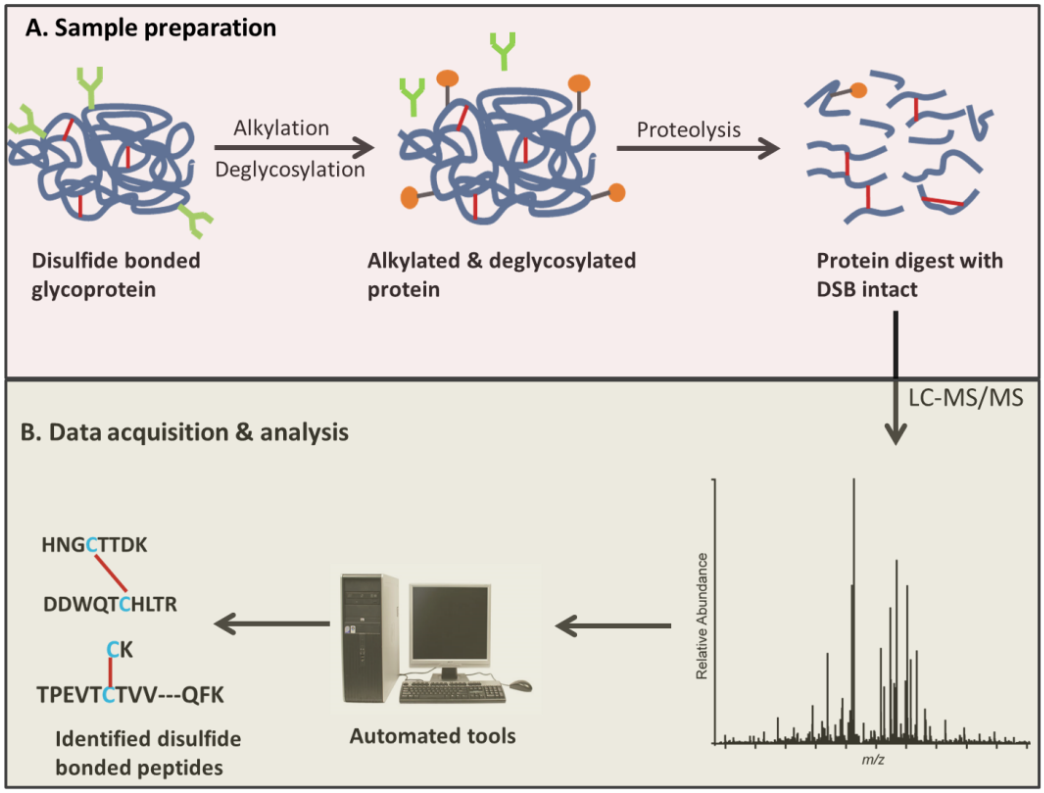


Figure 1. Disulfide bond analysis workflow. **(A.)** Sample preparation of non-reduced protein digest for disulfide bond analysis of a glycoprotein. **(B.)** Disulfide bond assignment from LC-MS/MS data. **Note:** deglycosylation is not necessary is the protein is not glycosylated.

1.3.2 Preventing Disulfide Bond Artifacts during Sample Preparation

A major requirement during sample preparation is to prevent the formation of non-native disulfide bonds (disulfide bond artifacts). To this end, a number of methods aimed at efficiently mapping disulfides in proteins without the formation of disulfide artifacts during sample preparation have been recently reported.[57,58] Disulfide bond artifacts can be induced during sample preparation via three main routes: (1) reaction between two free Cys residues ($R^1SH + R^2SH \rightarrow R^1SSR^2 + 2H$, free Cys oxidation), (2) reaction between a free Cys residue and a disulfide bond ($R^1SH + R^2SSR^3 \rightarrow R^1SSR^2 + R^3SH$, thiol-disulfide exchange), and (3) reaction between two Cys residues that were formerly involved in a disulfide bond ($R^1SSR^2 + R^3SSR^4 \rightarrow R^1SSR^3 + R^2SSR^4$, disulfide exchange).[59-62] These reactions are highly favored at alkaline pH[63,64] and high temperatures,[63,65,66] where disulfide scrambling usually occurs, especially for proteins containing free thiols. Besides pH and temperature, the thiol size, disulfide size, and the steric effects (exposure) of the thiols can also affect the formation of scrambled disulfide bonds, as shown by Kerr and coworkers[64]; but these factors cannot be controlled during sample preparation.

Temperature, pH, and the availability of free Cys are, therefore, the critical factors that must be controlled during sample preparation in order to prevent the formation of non-native disulfide bonds. Generally, samples for disulfide bond analysis are prepared at room temperature and then subjected to enzymatic digestion at 37°C, and such low temperatures do not trigger any cysteine reaction or disulfide bond shuffling.[63] However, pH tremendously affects disulfide bond or cysteine reactivity, even at room temperature, and must be carefully controlled during sample preparation.[63,64] Sample preparation should be done at slightly acidic pH because at alkaline pH,

free thiols are deprotonated, and the resulting thiolate anions are oxidized or react with adjacent disulfide bonds (thiol/disulfide exchange) to form new, non-native disulfide bonds. The thiol groups that are more exposed to solvent will be more reactive than those that are not exposed.[64]

Therefore, the first step during sample preparation for disulfide analysis is to cap any free cysteine residues in order to prevent formation of non-native disulfide bonds. This step is important even for proteins for which all the Cys residues are known to be in the disulfide bonded state because low levels of free Cys residues could be present. For example, although all the Cys residues in the four classes of IgG are expected to be disulfide bonded, low levels of free Cys residues have been detected in all 4 IgG classes,[33] and the free Cys residues can induce disulfide artifacts if they are not alkylated or not properly alkylated. Commonly used Cys alkylating reagents include: iodoacetamide (IAM), iodoacetic acid (IAA), and N-ethyl-maleimide (NEM).[67,68] Rogers and coworkers showed that NEM is a more suitable alkylating agent for protein thiols than IAM and IAA because it reacts faster, requires less reagent per mole of free Cys, and is very effective at acidic pH (pH 4.3 to 7.0),[68] while the other reagents are most effective at alkaline pH (pH 8.0), where thiol/disulfide exchange and free Cys oxidation reactions can compete with alkylation and lead to low levels of non-native disulfide bonds. Recently, Lu *et al* [69] reported low levels of disulfide artifacts when disulfide bond mapping of RNase A was done after alkylating free thiols with IAM and IAA at pH 6.5, but no artifacts were identified when the experiments were done at the same pH using NEM; indicating that IAM and IAA do not properly alkylate free thiols at slightly acidic pH conditions.[69]

1.3.3 Digestion

Protein digestion can be done either chemically or enzymatically.[70] Table 1 shows different types of disulfide-linked peptides that can result from protein digestion. Enzymatic digestion is the most widely used protein digestion approach, and there are a variety of enzymes to choose from, as reviewed by Switzer *et al*.[70] Selecting an appropriate enzyme for effective digestion is important because the enzyme used would determine the types of disulfide bonded peptides in the protein digest (Table 1). In selecting an enzyme, the goal is usually to choose one that produces more "simple" inter-chain disulfide-linked peptides, preferably with only one interchain disulfide bond and with bonded chains of appropriate lengths (4-15 AA residues). Such dipeptides can be easily mapped manually and with existing analysis software. For this purpose, *in silico* digestion of the protein using different enzymes is needed to select the best enzyme. Two important questions that can help in selecting a suitable enzyme are: (1) what are the sizes (number of disulfide bonded peptide chains and the length of each chain) of the disulfide linked peptides? and (2) How complex are the disulfide connectivity of the disulfide-linked peptides? That is, are there only interchain disulfide bonds, interchain and intrachain disulfide bonds, nested intrachain disulfide bonds? etc. If a single enzyme does not yield disulfide-bonded peptides with simple interchain disulfide bonds, a combination of enzymes can be used.[71,72] However, it is worth noting that using multiple enzymes could lead to very short disulfide-bonded peptides, which may not be retained in reverse-phase columns, making disulfide assignment difficult. In such cases, a planned digestion that deliberately ensures missed cleavages in order to obtain longer peptide chains can be done.[72] In addition to the type of disulfide bonded peptides generated, another important consideration is the digestion efficiency

of the enzyme used. Glatter *et al* studied several protein digestion strategies and showed that a combination of trypsin and Lys-C gives better digestion efficiency than trypsin alone.[73]

Table 1: Types of disulfide bonded peptides (DSBPs) from proteolytic digestions

No.	Protein	Enzyme(s)	Disulfide bonded peptides	Comment	Ref.
1	RNase A	Asp-N/C + trypsin	CKPVNTFVHESLA YPNCAYK	Two-chain DSBP with one interchain DSB	72
2	IgG3	Trypsin	SCDTPPPCPR SCDTPPPCPR	Two-chain DSBP with two interchain DSBs (disulfide box)	54
3	RNase A	Trypsin	CKPVNTFVHESLADVQAVCSQK YPNCAYK HIIVACEGNP- - -DASV	Three-chain DSBP with two interchain DSBs	72
4	HIV, gp140	Trypsin	NDMVDQMHEDIISLWDQSLKPCVK DNSDNSEYILINCDASTITQACPK LTPLCVTLHCDATFK DCSFDTTTEIR	Four-chain DSBP with three interchain DSBs - PNGase F used for deglycosylation	75
5	CTT, gelatinase inhibitor	N/A	CTTHWGFTLC	Completely cyclisized DSBP	91
6	Chicken lysozyme	Trypsin	NLCNIPCSALLSSDDITASVNCAK WWCNDDR	Two-chain DSBP with both interchain and intrachain DSBs	80
7	rhASA	Lys-C + trypsin + Asp-N	DQGPCQ DLTCFPPATPCDGGC	Two-chain DSBP with both interchain and intrachain DSBs	71
8	rhASA	Pepsin	PALQICCHPGCTPRPACCHCPDPHA	Single-chain DSBP with three intrachain DSBs (nested disulfides)	71

- Aspartic acids in green (D) represent asparagine residues that were converted to aspartic acids after deglycosylation with PNGase F.
- DSBP and DSB imply disulfide-bonded peptide and disulfide bond, respectively.

As mentioned earlier, the most important consideration (factor) during sample preparation for disulfide analysis is to prevent the formation of non-native disulfide bonds. Therefore, as with the case for alkylation, the pH and temperature are also critical during digestion. Proteolytic digestion is commonly done at 37°C, where no scrambling occurs. In fact, Wang *et al* recently showed that trypsin plus Lys-C digestion can also be done with high digestion efficiency at room temperature.[58] Hence, pH is, again, a critical factor in preventing disulfide shuffling during protein digestion. Since disulfide artifacts can be introduced when samples are prepared at alkaline pH,[59,60] protein digestions for disulfide bond mapping tend to be done at neutral or slightly acidic pH. Recently, Sung *et al* [57] investigated disulfide scrambling in lysozyme and bevacizumab (an IgG1 antibody drug) upon digestion with trypsin, trypsin plus Glu-C, and Lys-C at pH of 6 and 7; and thermolysin at pH ranging from 5 to 7. No disulfide scrambling was observed when digestion was done using trypsin, trypsin plus Glu-C, and Lys-C at pH of 6. However, disulfide scrambling was observed when digestion was done at pH 7 using the same enzymes, and at pH 5 to 7 using thermolysin.[57] Similarly, low levels of disulfide scrambling have also been reported when trypsin or Lys-C plus trypsin were used for protein digestion at pH of 6.8.[71,74,58] Nonetheless, disulfide scrambling at neutral or slightly acidic pH conditions may depend on a variety of conditions in addition to pH, since no scrambling was observed when disulfide mapping of bovine fetuin was done after trypsin digestion in buffers at pH of 5.5, 6.5, and 7.[75] In some cases, disulfide scrambling at slightly acidic pH may result from lack of effective alkylation of any free Cys residues prior to digestion. One useful strategy in determining whether or not the particular digestion conditions to be used introduce artifacts is to check the sample preparation method using a standard protein whose disulfide bonding is known,

prior to analyzing an unknown protein; this strategy has been demonstrated previously.[75] It is worth noting that trypsin and Lys-C digestion at slightly acidic pH can lead to missed cleavages due to incomplete digestion, as the optimal efficiency of the enzymes are at pH 8.0 and pH range of 8.0 to 8.8, respectively;[70] therefore, a single set of digestion conditions is not necessarily optimal for every protein.

In order to completely eliminate the possibility of disulfide scrambling, digestion can be done using pepsin, which efficiently digests proteins at highly acidic pH (pH <2),[71,74,76] where the formation of non-native disulfides via free Cys reactivity is not possible. However, pepsin is less specific than trypsin or Lys-C, and it may produce very small disulfide-bonded peptides that would be difficult to separate and analyze. Nonetheless, the use of separate pepsin and trypsin (pH 6) digestions could be helpful for unambiguous assignment of disulfide bonds in cases where ambiguity in the disulfide bond pattern has been a problem.[76]

1.3.4 Sample Preparation of Glycoproteins

Disulfide bond analysis of glycoproteins may require deglycosylation prior to mass spectrometric analysis. In order to decide whether or not a glycoprotein needs to be deglycosylated, the occupied glycosylation sites of the protein must first be identified, and *in silico* digestion using an enzyme of choice (or a combination of enzymes) can be done to decipher whether or not the Cys-containing peptides that are disulfide-linked would contain occupied glycosylation sites. For some glycoproteins (e.g monoclonal antibodies) deglycosylation is not necessary because the disulfide-bonded peptides resulting from digestion using the commonly used enzymes usually do not contain a glycosylation site.[65] However, other glycoproteins may contain one or more occupied glycosylation sites near cysteine residues

involved in a disulfide bond, and in such cases, deglycosylation must be done in order to reduce the complexity of the MS/MS data of the disulfide bonded peptides, because glycosylation cannot be simply regarded as a modification on the peptide chains.[71,75,77] For example, trypsin digestion of the HIV-1 Env sequence variant, C97ZA012 gp140, which contains 25 N-linked glycosylation sites and 10 disulfide bonds, resulted in seven tryptic, disulfide-liked peptides that all contain at least one occupied glycosylation site on the disulfide-linked chains.[75] One of the tryptic digests was a four chain disulfide-linked peptide with six occupied glycosylation sites. In cases of this complexity, deglycosylation is a necessary first step.

Deglycosylation is typically done using peptide N-glycosidase F (PNGase F) at neutral or slightly acidic pH,[71,75,78] and the reaction can extend for several days if the protein is heavily glycosylated.[75,78] Hence, proper Cys alkylation must be done to prevent disulfide shuffling during such long incubation periods. In addition, deglycosylation converts the occupied asparagine residue (N) to an aspartic acid residue (D), leading to a 0.985 Da mass change, which must be considered when calculating the molecular masses of the deglycosylated disulfide-linked peptides.

1.4 Liquid Chromatography-Mass Spectrometric Disulfide Bond Analysis of Proteins

1.4.1 Separation of Protein Digests

Disulfide bond mapping by liquid chromatography coupled to mass spectrometry requires the separation and ionization of protein digests prior to mass spectrometric analysis. The widely used separation technique is reverse phase liquid chromatography (RPLC) by means of columns packed with either C_8 or C_{18} stationary phases and mobile phases consisting of polar (e.g water) and non-polar (e.g acetonitrile) solvents containing modifiers such as formic acid and

trifluoroacetic acid (0.01 to 0.1%, v/v). Different types of columns can be used for peptide separation, including microbore (e.g 1.0 and 2.1 mm internal diameter, i.d),[54,79] capillary (e.g 0.5 mm i.d) [80] and nano columns (typically 0.075 mm i.d).[65] The typical particle size is between 3 to 5 µm. Columns packed with 1.7 µm particles are becoming more common in recent years, but they require much higher pressures for separation.

Peptide separation is usually followed by online electrospray ionization and tandem mass spectrometry (ESI-MS/MS) analysis. However, in some cases where the disulfide-linked peptides to be analyzed are very large and in low abundance, offline LC fractionation can be used to collect and concentrate fractions containing the disulfide-linked peptides (and carryout further digestion if necessary), prior to MS/MS analysis.[5]

1.4.2 Fragmentation Techniques for Disulfide Bond Analysis

Collision induced dissociation (CID), a dissociation method that reacts the ion of interest with neutral gas molecules, and electron transfer dissociation (ETD), a dissociation method that reacts the ion with an electron carrier, are the most commonly used fragmentation techniques for mass spectrometry-based disulfide bond analysis.[54,80,81] Figure 2 shows CID and ETD spectra of a simple interchain disulfide-bonded peptide from IgG3 monoclonal antibody that exhibits the characteristic fragment ion peaks that result from subjecting disulfide-bonded peptides to CID (Figure 2A) and ETD (Figure 2B) fragmentation.

CID typically leads to the fragmentation of the peptide backbone (amide) bonds, while leaving the disulfide bond intact; hence producing b and y ions that contain the disulfide bond (ions in red in Figure 2A), as well as ions that do not contain the disulfide bond (ions in green, Figure 2A). Although assignment of CID-generated product ions of disulfide-bonded peptides

was previously based on the assumption that only one peptide bond is cleaved during CID, Clark *et al* [80] recently showed that cleavage of two peptide bonds (double cleavage) is common during CID fragmentation of disulfide-linked peptides. The peaks labeled in red brackets in Figure 2A are examples of this type of cleavage. Although CID does not typically cleave disulfide bonds, a few instances have been reported where thioaldehydes, persulfides, and dehydroalanine ion peaks resulting from cleavage of the C-S and S-S (disulfide) bonds are present in CID spectra.[72,82]

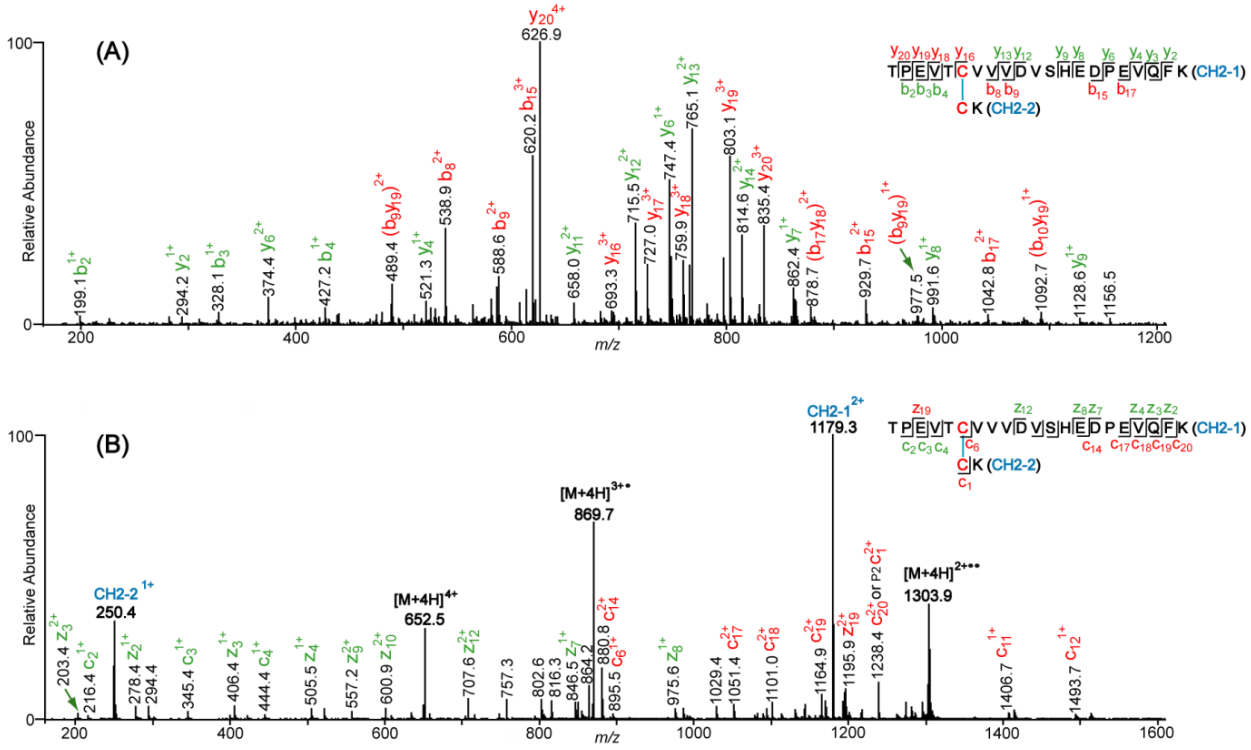


Figure 2. CID **(A)** and ETD **(B)** spectra of a disulfide bonded peptide from the CH2 domain of IgG3 monoclonal antibody. The spectra show the characteristic b/y **(A)** and c/z **(B)** ions resulting from the backbone cleavage of the peptides linked by the disulfide bond. Fragment ions not containing the disulfide bond are labeled in green and those containing the intact disulfide bond are labeled in red. CID fragment ions resulting from double cleavage and containing the intact disulfide bond are in brackets. In the ETD spectrum **(B)**, intense peaks of the Cys-containing peptides linked by a disulfide bond are observed at m/z 250.4 and 1179.3. These peaks result from the cleavage of the disulfide bond by ETD, and they are not observed in **(A)**.

In contrast to the product ions generated from CID of disulfide bonded peptides, cleavage of the disulfide bond is the primary reaction pathway during ETD fragmentation. Backbone (amide) cleavage also occurs, by generation of c and z ions, although this fragmentation pathway generally occurs at a lesser extent.[83,84] As a result, fragment peaks representing the disulfide-linked peptides (e.g peaks at *m/z* 250.4 and 1179.3 in Figure 2B) are usually more intense in ETD spectra than c/z ion peaks.[53,54,83] Peaks resulting from backbone fragmentation of the bonded peptides may or may not contain the intact disulfide bond (fragment ions labeled in red and green in Figure 2B).

Besides CID and ETD, higher-energy collisional dissociation (HCD)[85] and a dual fragmentation technique known as electron-transfer and higher-energy collision dissociation (EThcD)[86,87] are newer fragmentation techniques available in some Orbitrap mass spectrometers, and they are gaining ground in disulfide bond analysis.[79,69,74,] HCD has similar fragmentation pattern to beam-type CID,[88] and it produces only b/y ions, while EThcD spectra exhibit product ions present in both ETD and CID spectra (containing b/y and c/z ions), as well as ions resulting from disulfide bond cleavage.[74] In addition, ultraviolet photodissociation (UVPD) is another fragmentation technique that is being applied for disulfide bond analysis.[89,90] Fragmentation of disulfide-linked peptides by UVPD leads to selective scission of the disulfide bond, as observed under ETD conditions.[90]

1.4.2.1 Considerations for Selecting Fragmentation Techniques for Disulfide Analysis

Choosing a suitable fragmentation technique or combination of techniques for data acquisition is vital in facilitating disulfide bond assignments from MS/MS data. The choice of a fragmentation method is typically based on the complexity of the disulfide linkages, the size of

the disulfide bonded peptides, and, to a lesser extent, on the amino acid sequence of the bonded chains. Although either CID or ETD can be used to assign simple interchain disulfide bonded peptides (as shown in Figure 2), a careful choice of a fragmentation method is necessary for the assignment of complex disulfide linkages involving either intrachain or both interchain and intrachain disulfide bonds, large disulfide-bonded peptides with multiple peptide chains, as well as very small disulfide bonded peptides.

ETD is preferred for the fragmentation of peptides that are completely cyclized by a disulfide bond, as recently demonstrated by Xia and colleagues.[91] CID fragmentation of such peptides would typically not reveal any backbone sequence ions, whereas ETD leads to cleavage of the disulfide bond and the peptide backbone, thereby revealing ions that can be used to assign the disulfide bond.[91] ETD followed by a CID-MS^3 step (ETD-CID-MS^3), as reported by Karger *et al*,[52] as well as a fragmentation technique involving hydroxyl radical addition followed by CID fragmentation, as reported by Durand *et al*,[92] can also be used to map such disulfides.

For complex disulfide cases, such as nested disulfides and cysteine knots, neither ETD nor CID may be sufficient as a standalone technique to unambiguously assign the complex disulfide bond connectivity. In such cases, methods involving combined ETD and CID fragmentations would be ideal to decipher the disulfide connectivity.[71,93] The EThcD fragmentation method could be promising in mapping these complex disulfides, but it has not yet been applied to such cases.

In terms of size, ETD is preferable for the assignment of large disulfide-bonded peptides containing three or more peptides linked by interchain disulfide bonds. Such large peptides can

easily ionize in sufficiently high charge states for ETD analysis, since ETD requires highly charged ions for efficient fragmentation. Due to preferential cleavage of the disulfide bonds, ETD spectra of large multi-chain DSBPs are less complex, and they clearly reveal which peptides are directly linked to each other.[75] However, CID spectra of such large disulfide-linked peptides are complex and difficult to assign (e.g because of several double cleavages), and the spectra do not reveal which of the peptide chains are directly bonded to each other. For example, Go *et al* [75] assigned a three-chain disulfide-linked peptide using both CID and ETD spectra, but only ETD was used to unambiguously assign a four-chain disulfide-linked peptide. The ETD spectra of the three-chain and the four-chain disulfide-bonded peptides clearly indicated the chains that were disulfide-linked to each other.[75]

On the contrary, CID is ideal for analysis of small disulfide-bonded peptides and disulfide-linked peptides with peptide chains that contain several adjacent proline residues. CID is preferable in such cases because small disulfide bonded peptides may not ionize at high charge states, which are required for ETD analysis, and peptide chains containing adjacent proline residues fragment efficiently when subjected to CID but not ETD fragmentation, due to the N-C_α ring structure of proline.[94] For example, the hinge region of IgG3 contains a small dipeptide, CPEPK linked to CPEPK, and SCDTPPPCPR disulfide-linked to SCDTPPPCPR and a recent analysis of IgG3 disulfide bonds showed that these peptide chains did not fragment well under ETD because of the small size (for the first disulfide-linked peptide) and several adjacent proline residues on the peptide chains (for both disulfide-linked peptides), but they were unambiguously assigned using CID data.[54] Hence, in addition to the size of a disulfide-linked peptide, the amino

acid sequences of the Cys-containing peptides could also be a significant factor in selecting an efficient fragmentation technique for disulfide analysis.

Given the above explanations, it is imperative for researchers to do *in silico* digestion of a protein in order to study the possible types of disulfide linked peptides (see Table 1) and to make a decision on which fragmentation method (or combination of methods) would be suitable for unambiguous mapping of the disulfide bonds in a particular protein. In general, the collection of both ETD and CID data proves to be very useful, since the data sets are so complementary.[53,54]

1.4.3 Bottom-up Methods for Disulfide Bond Analysis

A variety of bottom-up mass spectrometric methods based on the afore-mentioned fragmentation techniques have been reported in recent years. Figure 3 shows a schematic representation of bottom-up approaches for disulfide bond analysis. The methods generally fall under two broad categories; disulfide bond mapping from: (1) reduced and non-reduced protein digests (profile comparison), and (2) only intact (non-reduced) protein digests. The second category can be further divided into methods that require chemical or electrochemical post-column reduction of disulfide bonds, reduction of the disulfide bonds in the gas phase (ETD and EThcD fragmentations), and those that do not require any reduction (typically CID and HCD methods). In the following sections, we review the bottom-up methods that have been developed in the past decade.

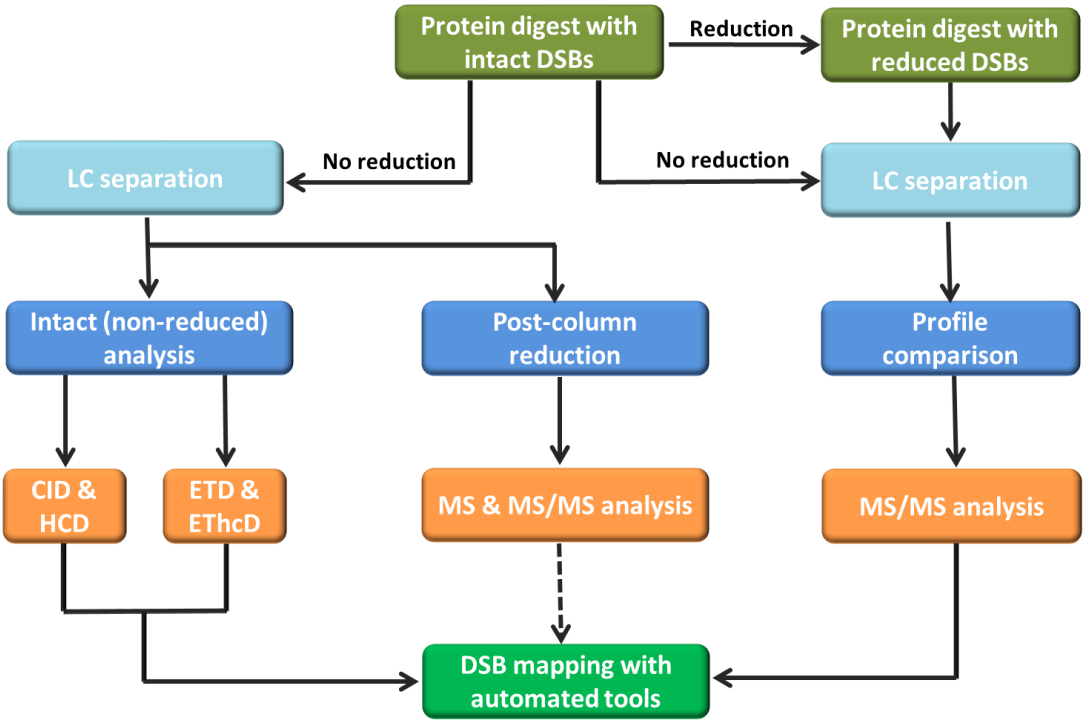


Figure 3. Schematic representation of bottom-up approaches for disulfide bond analysis. The methods generally fall under two broad categories: (1) disulfide bond mapping from protein digests with reduced and non-reduced disulfide bonds (profile comparison), and (2) disulfide bond mapping from only intact (non-reduced) protein digests. Non-reduced analysis can be further divided into methods that involve post-column reduction of disulfide bonds, reduction of the disulfide bonds in the gas phase (ETD and EThcD fragmentation methods), and those that do not require reduction of the disulfide bonds (CID and HCD fragmentation methods). Data analysis after post-column reduction has not been automated yet.

1.4.3.1 Profile Comparison: Reduced and Non-reduced Analysis

Disulfide bond mapping by profile comparison is usually done using two samples: a protein digest without reduced disulfide bonds (non-reduced sample) and a digest with reduced disulfide bonds (reduced sample). The two samples are separated in two LC runs (experiments), and peptides that are present in the ultraviolet (UV) or total ion chromatogram (TIC) profile of

the non-reduced sample, but are absent in that of the reduced sample, are generally considered to be disulfide bonded, and MS/MS analysis can be done to assign the disulfide bonds.[55,56,77]

Herein, we present two recent examples where complete or partial reduction of the disulfide bonds was used to assist in disulfide mapping. Researchers at Genentech used the reduced and non-reduced peptide mapping technique to assign the disulfide bonds of an unusual IgG1 antibody variant that was identified after size exclusion chromatography (SEC).[95] Results from several experiments suggested that the mAb variant contained an extra light chain connected to the IgG1 monomer via a disulfide bond. LC-MS/MS experiments using non-reduced and reduced digests of the variant were used to confirm the disulfide linkage site of the extra light chain to the correct IgG1 monomer. In a second related example, an approach that deviates from the standard profile comparison method was reported by Klapoetke *et al.*[96] In order to alleviate the need for preparing and analyzing two separate samples, these researchers conducted a partial reduction of the disulfide bonds and alkylated the resulting resulting free Cys, prior to enzymatic digestion. LC-MS/MS analysis of the mixed sample, containing partially reduced and non-reduced disulfide-linked peptides, was used to assign the disulfide bonds of the protein. The method is suitable to assign complex disulfides such as Cys knots, while simple disulfides can be mapped using a non-reduced sample.[96] The challenging aspect of this approach is to optimize the reducing agents to obtain partial reduction, as well as optimizing the alkylating agent to completely alkylate the reduced species.

Overall, a shortcoming of the techniques using disulfide reduction in the workflow is the need to prepare and analyze two samples, or to work out sample prep conditions that afford

partial reduction, thereby increasing sample preparation and analysis time. In addition, the method is suitable only for disulfide bonds containing only interchain disulfides.

1.4.3.2 Intact (Non-reduced) Analysis

An alternative approach to mapping disulfide bonds is by peptide mapping of non-reduced protein digests (intact digests) without the need for reduced aliquots. The methods for disulfide bond assignment from non-reduced protein digests can be divided into three categories: methods involving post-column chemical or electrolytic reduction of the disulfide bonds, those involving gas-phase reduction of the disulfide bonds, and methods that do not involve cleavage of the disulfide bonds. In this section, we describe the methods that fall in the first two categories. For the last category, the disulfide bonds are typically assigned by peptide mapping using CID data by calculating theoretical *m/z'*s (at different charge states) of the expected disulfide bonded peptides and searching the calculated *m/z* values in the MS data; the ions are assigned by identifying the characteristic fragment ions in the MS/MS spectra.[78,80] Several automated tools have been developed to identify disulfide bonded peptides without cleaving the disulfide bond, and those are discussed in the automated tools section (Section 1.6).

1.4.3.2.1 Post-Column Reduction Methods

Several methods have been developed whereby non-reduced protein digests are separated by liquid chromatography followed by post-column partial reduction of disulfide bonded peptides prior to mass spectrometric characterization, typically with CID fragmentation. Post-column reduction of disulfide bonds can be done chemically by introducing a reducing agent such as Tris (2-carboxyethyl) phosphine (TCEP), by in-source reduction during ionization, or

electrochemically via an increased potential on an electrochemical cell placed on the flow path from the column.

Post-column chemical reduction of disulfide-bonded proteolytic digests typically involves separation of the non-reduced protein digests by reversed phase liquid chromatography, and the eluates from the column are mixed (via a mixing-tee) with an optimized amount of a reducing agent (e.g TCEP) to partially reduce the disulfide bonds, as demonstrated by Li *et al.*[97,98] Since only partial disulfide bond reduction is done after LC separation, disulfide-linked peptides with intact disulfide bonds and their corresponding Cys-containing peptides that resulted from the disulfide bond reduction would have the same retention time and would both be detected in the same full scan mass spectrum. To identify disulfide-bonded peptides, the extracted ion chromatograms of all Cys-containing peptides are compared, and the Cys-containing peptides that have the same retention time are assigned as disulfide bonded.[97,98] The assignments are confirmed by the presence of a disulfide-linked peptide at the same retention time with a molecular mass corresponding to the sum of the molecular masses of the two Cys-containing peptides minus 2 Da. Scrambled disulfide bonds are readily assigned in the same manner.[98] Liu and coworkers[99] used this approach to map the disulfide bonds of an IgG2 antibody, but instead of using only LC-MS data, LC-MS/MS (CID) data were used for unambiguous characterization of the LC peaks of the disulfide bonded peptides and those of their corresponding Cys-containing peptides.

Besides chemical reduction, in-source reduction during electrospray ionization can be used for partial reduction of disulfide-linked peptides, as shown by Cramer *et al.*[79] Post-column in-source reduction, followed by HCD analysis of the partially reduced peptides, was used to

develop a method capable of assigning complex intra- and inter-chain disulfide bonds in proteins, including disulfide bonds involving closely-spaced Cys residues.[79]

Methods involving post-column electrochemical (EC) reduction of disulfide bonds prior to mass spectrometric analysis are also emerging approaches for disulfide bond characterization in proteins. Cramer *et al* [100] recently demonstrated that disulfide bond assignment in small proteins, such as human insulin, can be done by direct infusion of the intact protein into an electrochemical cell for partial reduction of DSBs of the intact proteins followed by tandem mass spectrometry (CID fragmentation) to confirm the sequence coverage of the protein. For larger proteins (e.g human serum albumin and ribonuclease B), proteolytic digestion and LC separation of the protein digests are necessary prior to post-column EC reduction of the disulfide-bonded peptides and MS/MS fragmentation (LC-EC-MS/MS), as recently demonstrated by Switzar *et al*.[101] A rapid method employing on-column proteolytic digestion and electrochemical reduction was reported by Chen *et al*.[102,103] In this approach, an online pepsin column is used for rapid digestion of intact proteins, and disulfide bonded peptides are partially or completely reduced using different electrochemical potentials, followed by desorption electrospray tandem mass spectrometry (DESI-MS/MS).[102] This online digestion approach provides unprecedented digestion and reduction speeds (done in less than 10 mins) and very low potential for inducing disulfide artifacts due to less sample handling and pepsin digestion at low pH.

In general, although the chemical and electrochemical post column partial reduction methods do save sample preparation time (compared to the case of reduced and non-reduced samples), optimizing the reduction conditions to get sufficient signals for both the reduced and non-reduced peptides is a major challenge.

1.4.3.2.2 Gas Phase Dissociation (Reduction) Methods

Because ETD mainly cleaves disulfide bonds (gas phase reduction) leading to intense Cys-containing peptide product ions, several disulfide bond characterization methods have been developed that take advantage of the unique ETD fragmentation pattern to analyze non-reduced protein digests for disulfide bond mapping. Karger *et al* reported a method that combines alternating ETD/CID fragmentation (MS^2) of the disulfide-bonded peptides and CID-MS^3 fragmentation of the Cys-containing fragment ions that resulted from the ETD-MS^2 dissociation of disulfide bonds.[104] This method is suitable to map disulfide-bonded peptides containing complex disulfide linkages, such as nested disulfides and cysteine knots,[71] as well as peptides completely cyclized by a disulfide bond.[91] However, due to the enormous data generated as a result of the additional CID^3 step, the method may not be necessary for samples that do not contain complex disulfides. Recently, Massonnet and coworkers[93] reported a similar technique for disulfide bond assignment of peptides containing two disulfide bonds. After opening of the disulfide bonds by ETD, the generated species are separated by ion mobility prior to characterization of the separated ions by CID fragmentation.[93]

Clark *et al* reported a much simpler ETD-based method that is suitable for samples containing simple interchain disulfide-bonded peptides, including disulfides with multiple (three to four) disulfide bonded chains.[53] The method does not require an MS^3 step, and instead of searching the *m/z*'s of all expected (known) disulfides and possible scrambled disulfides, the extracted ion chromatograms (XIC's) of all Cys-containing peptides are generated, and the ETD spectra of the peaks in the XICs are interrogated to determine the bonding partner(s) of the Cys-containing peptides.[53,54] Disulfide assignments are made when fragment ions of Cys-containing

peptides are identified in the same ETD spectrum, and the sum of their masses minus 2 Da (mass of 2H) match the mass of the expected disulfide. The *c* and *z* ions resulting from backbone cleavage of the bonded peptides can be used to further confirm the identities of the bonded chains.[54] This XIC approach is suitable to map the disulfide bonds in proteins whose disulfide connectivity is unknown and also to quickly verify non-native or alternative disulfide bonds. In addition, the XICs of Cys-containing peptides can also provide information about any free Cys in the protein without the need for differential alkylation.[105,106]

Similar to ETD fragmentation, ultraviolet photodissociation (UVPD) of simple interchain disulfide bonded peptides in the gas phase can be used to map disulfide bonds in proteins, as demonstrated by Agarwal *et al*.[90] After separation of non-reduced protein digests using LC, irradiation of disulfide-linked peptides, including three chain disulfide-linked peptides, with laser pulses at 266 nm in an ion-trap mass analyzer led to selective cleavage of disulfide bonds, thus revealing the disulfide bonded peptides. However, backbone cleavage of the Cys-containing peptides is not typically observed using this approach.[90]

1.5 Software for Automated Disulfide Bond Assignment

Manual analysis of disulfide bonds in proteins is a challenging and time-consuming task that requires expertise in the field. As such, an increasing number of automated tools have been developed for rapid and high-throughput protein disulfide bond assignment using MS/MS data. Automated tools are ugently needed when large cohorts of samples need to be routinely analyzed, as is the case in biopharmaceutical industries. Most of the developed tools are based on the non-reduced analysis, although a few tools also have the option to map disulfide bonds by comparing data from reduced and non-reduced samples.

MassMatrix, developed about a decade ago, is one of the prominent tools for automated assignment of disulfide-linked peptides from non-reduced protein digests.[107] Assignments are solely based on fragment ions from the backbone cleavage of the bonded peptides. This is a shortcoming of the tool, because only a limited number of ions are considered for disulfide bond assignment. Most importantly, fragment ions containing the disulfide bonds are not considered for disulfide bond mapping. Several tools have since been developed which address this problem. Dbond [72,82] and MS2DB+[108,109] can map disulfides from MS/MS data using several characteristic fragmentation peaks of disulfide bonded peptides, including b/y or c/z ions containing intact disulfide bonds. While DBond assigns dipeptides solely from CID data, and includes ions resulting from the occasional CID cleavage of C-S and S-S bonds (resulting in thioaldehydes, persulfides and dehydroalanin ions),[72,82] the MS2DB+ software can assign disulfides from either CID or ETD data, and the fragment ions considered for CID data include ions resulting from the loss of water or ammonia from b/y ions. For both software products, the theoretical *m/z*'s of all possible disulfide-bonded peptides are searched from MS^1 data, and their identified precursor ions are automatically assigned by comparing theoretical and experimental MS/MS spectra of the disulfide-bonded peptides. In addition, the tools score the assigned disulfide-bonded peptides based on the number of matched fragment ions and their intensities.

A unique algorithm known as RADAR (Rapid Assignment of Disulfide Linkage via A1 ion Recognition) that assigns disulfide bonded peptides based on dimethyl labeling of the N-termini of the bonded chains was reported by Huang *et al*,[110] and it has been applied for disulfide bond assignment in monoclonal antibodies.[111] Dimethyl-labeled peptides produce intense a1 ion signals in CID spectra, which are used to screen for disulfide bonded peptides. Because these

species mostly contain two or more peptide chains, the CID spectra would have two or more a1 ion peaks, and RADAR rapidly screens for disulfide-linked peptides by first searching for such spectra that contain multiple a1 ion peaks.[110] Cys-containing peptides whose a1 ions are found in the same spectrum are considered to be disulfide bonded, and they are confirmed by matching the precursor ion mass to the molecular weight of the predicted disulfide-linked peptide, followed by assigning the fragment ions (b/y ions) resulting from fragmentation of the disulfide-bonded peptide chains.[111] This approach is suitable for rapid identification of disulfides, including scrambled disulfides, because prior knowledge of the disulfide connectivity is not required. Intrachain disulfides on a single peptide can be assigned in a similar manner but instead of searching for multiple a1 ions in the CID spectra, a specific (target) a1 ion corresponding to the N-terminal of the peptide can be searched for. The software is effective for disulfide mapping in complex protein mixtures (containing more than 20 disulfide bonds) and for rapid verification of disulfide scrambling.[112]

Tools that assign disulfide bonds from HCD and EThcD data have also been developed within the last decade. A software tool known as pLink-SS was introduced in 2015 by Lu and coworkers for automated disulfide bond assignment in proteins using HCD data.[69] The software was used to map simple and complex disulfides from a purified protein (IgG), a mixture of 10 proteins containing 74 disulfide bonds, and at the proteome level (199 DSBs in 150 proteins from *E. Coli*) using HCD data.[69] Another tool known as SlinkS assigns interchain and intrachain disulfide bonds by matching experimental MS/MS fragment ions from ETD and EThcD spectra to *in silico* fragments.[74] For assignments from ETD data, the software searches for marker ions corresponding to the masses of the disulfide-linked peptides, as well as c/z ions with and without

disulfide bonds; while b/y ions are also included in the search when assigning disulfide bonds from EThcD data. The assigned disulfide-linked peptides are scored, and FDRs are calculated based on user-defined cut-off values.

Besides the aforementioned tools, several commercial software products are available for automated analysis of protein disulfide bonds connectivity using MS/MS data, particularly in protein pharmaceuticals. These include BioConfirm (Agilent Technologies), Byologic (Protein Metrics Inc), DisulfideDetect (Bruker), PepFinder, BioPharma Finder, and Pinpoint (ThermoFisher Scientific), BioPharmView (SCIEX), and BiopharmaLynx (Waters). While tools like DisulfideDetect require only non-reduced protein digest for disulfide analysis, PepFinder and BioPharma Finder can map disulfide-bonded peptides using either non-reduced protein digests or both reduced and non-reduced protein digests.

Although the automated assignment of disulfide bonded peptides has facilitated disulfide analysis of proteins, there is still room for improvement in this field. Most of the tools developed so far can readily assign simple inter-chain disulfide-linked peptides but would not readily assign disulfide-bonded peptides containing complex intertwined disulfide linkages involving both interchain and intrachain disulfide bonds, nested disulfides, and dipeptides with more than one inter-chain disulfide bond (disulfide box). Although several approaches can be utilized to manually assign the disulfide bonds of these complex cases,[113] software that can readily assign such disulfide-bonded peptides remain elusive.

1.6 Challenges and Summary

1.6.1 Challenges in Disulfide Bond Analysis

Despite the existence of high resolution mass spectrometers, automated tools, and several strategies to map disulfide bonds in proteins, there are still some challenges that can be encountered during disulfide bond assignments. When non-native disulfide bonds are identified after bottom-up disulfide analysis, it is difficult to tell if they are disulfide artifacts introduced during sample preparation or they are alternative disulfide bonds (disulfide variants) present in the sample. One way to verify if disulfide artifacts are introduced during sample preparation is to validate a particular method using a protein whose disulfide bond connectivity has been well characterized.[78] In addition, top-down analysis of the intact protein requires less sample handling, and it would be an important first step in pinpointing the source of non-native disulfide bonds.[114] Finally, when proteins have been identified to contain alternative disulfide bonds (disulfide-mediated isoforms), quantifying the amount of the protein that contains the native and alternative disulfide bonds remains a challenge because, many times, the disulfide-mediated isoforms would not be easily separated by standard separation techniques. State-of-the-art separation methods are a possible route forward for solving this problem, though baseline separation of the isoforms remains a challenge.[115,116]

1.6.2 Summary of Disulfide Bond Analysis

Disulfide bond mapping in proteins provides vital information regarding the disulfide bond connectivity pattern in the protein, and it is thus important to confirm the global disulfide bond structure of a particular protein and to verify whether or not any disulfide-mediated isoforms[5] or variants[95] are present. Several bottom-up mass spectrometry approaches, as

described herein, have been developed for this purpose. *In silico* digestion is an important first step for disulfide mapping to decide what enzyme (or combination of enzymes) to use for digestion in order to get simple disulfide bonded peptides that are easy to analyze, and to decide what fragmentation technique would be suitable for easy and rapid assignment of disulfide bonded spectra. The majority of disulfide mapping methods are based on CID and ETD fragmentation techniques, but with newer fragmentation techniques such as EThcD and HCD, new disulfide mapping methods involving these techniques will be developed in the future. From a data analysis point of view, software tools will need to be developed for automated characterization of disulfide-bonded peptides with complex disulfide bond connectivity. Finally, quantitation of disulfide variants remains a challenge.

1.7 Protein Glycosylation

1.7.1 Overview of Protein Glycosylation

Protein glycosylation, the covalent attachment of oligosaccharide chains to proteins, is a complex post-translational modification process that occurs in the endoplasmic reticulum (ER) and Golgi apparatus during protein synthesis in cells.[117] This process of conjugating oligosaccharide chains (glycans) on proteins results in the formation of glycoproteins, and more than 50% of all proteins are glycosylated.[118] There are different types of protein glycosylation, including *N*-linked,[119] *O*-linked,[120, 121] *C*-linked,[122] and *S*-linked[123] glycosylation; with *N*- and *O*-linked glycosylation being the most common types. *N*-linked glycosylation results from the attachment of *N*-glycans to a nitrogen atom on the side-chain of asparagine (N) amino acid residue while *O*-linked glycosylation results from the attachment of *O*-linked glycans to a hydroxyl group on the side chains of serine (S/Ser) or threonine (T/Thr) amino acid residues.

While the formation O-linked glycosylation only requires the presence of Ser and/or Thr amino acid residues, *N*-linked glycosylation requires the asparagine (N) residue to be within a consensus amino acid motif of N-X-S/T, where X could be any amino acid except proline.[117] However, not all asparagine residues in the glycosylation site motifs are glycosylated.

Glycosylation plays critical structural and functional roles in glycoproteins, and it is also important during the development of glycoprotein-based biotherapeutics. Glycosylation can impact a protein's folding, half-life, transportation, cellular interactions, as well as solubility.[124-126] In addition, changes in the glycosylation levels of endogenous proteins can provide information relating to disease progression.[127] In recent years, several glycoproteins, especially monoclonal antibodies, have been used for the development of biotherapeutics to treat diseases.[128] The glycosylation profiles of these drugs are important for cell line selection, optimization of cell culture process parameters during the recombinant synthesis of the proteins, ensuring batch-to-batch consistency during drug development, and for comparing biosimilars to innovator biologics.[129-132] Glycosylation is also a major critical quality attribute of biologics that must be within a certain pre-determined range in order to ensure safety and efficacy of the drugs. Hence, identifying and characterizing the types of glycans attached to proteins is vital for effective development of efficacious and safe biotherapeutics and vaccines, for structural and functional studies of glycoproteins, as well as biomarker discovery for disease diagnosis.

The attachment of glycans to proteins imparts tremendous complexity and diversity on the proteins due to the variety of *N*- and *O*-linked glycans that could be appended to the proteins.[133, 134] Different types of glycans can occupy different glycosylation sites (macroheterogeneity) and/or the same glycosylation site (microheterogeneity). This macro and

micro glycosylation heterogeneity further increase the complexity of glycoproteins and presents an enormous challenge in protein glycosylation profiling. Figure 4 shows some common *N*-glycans that can be attached to proteins. Each of the glycans consists of a conserved core of two *N*-acetylglucosamine (GlcNAc) and three mannose (Man) residues. *N*-linked glycans fall into three main categories: high-mannose, hybrid, and complex glycans; and all three *N*-glycan types are formed from a precursor glycan ($Glc_3Man_9GlcNAc_2$) via several enzymatic processes.[117] Several *O*-linked glycans have also been reported, with diverse core structures.[134]

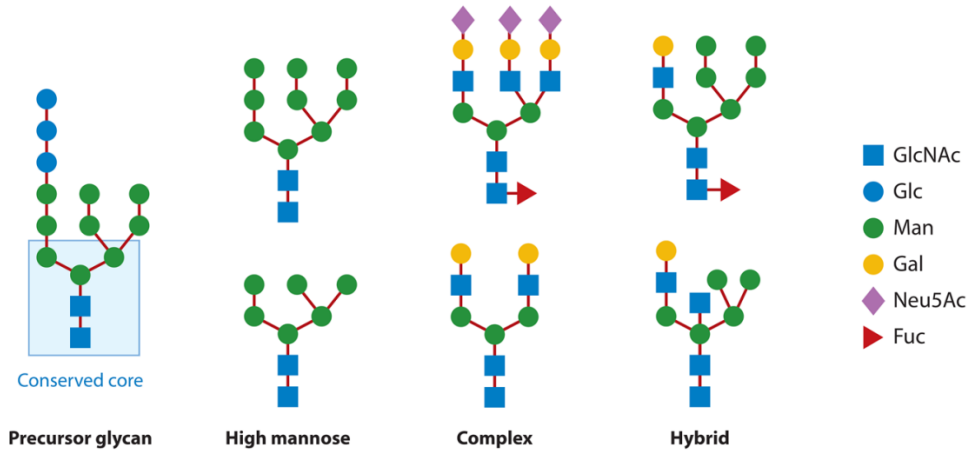


Figure 4. Examples of *N*-linked glycan structures. The monosaccharide residues include N-acetylglucosamine (GlcNAc), glucose (Glc), mannose (Man), galactose (Gal), N-acetylneuramic acid (Neu5Ac), and fucose (Fuc). This figure is reproduced from Reference 133.

1.7.2 Characterization of Protein Glycosylation

Two main methods are commonly used for glycosylation profiling: Glycan analysis and glycopeptide analysis. Figure 5 illustrates the workflow for glycosylation analysis using these two approaches. Glycan characterization (Fingure 5A) requires the release of the glycans from the protein prior to analysis, typically by liquid chromatography and fluorescence and/or a mass spectrometric detection (LC-FLD or LC-MS/MS).[135-137] *N*-glycans are usually released from glycoproteins using Peptide-*N*-Glycosidase F (PNGase F) enzyme. All glycans are naturally not fluorescent; hence, glycan analysis by LC-FLD requires derivatization of the released glycans with fluorescence tags (e.g 2-amino benzoic acid and *RapiFluor*) in order to enable fluorescence detection. The released and derivatized glycans are typically purified from the rest of the protein using solid phase extraction, and the purified glycans can be separated using hydrophilic interaction chromatography (HILIC).[135] The glycan analysis approach provides information about the total glycan types present on a protein but glycosylation site-specific information is lost, except for proteins with only one glycosylation site.

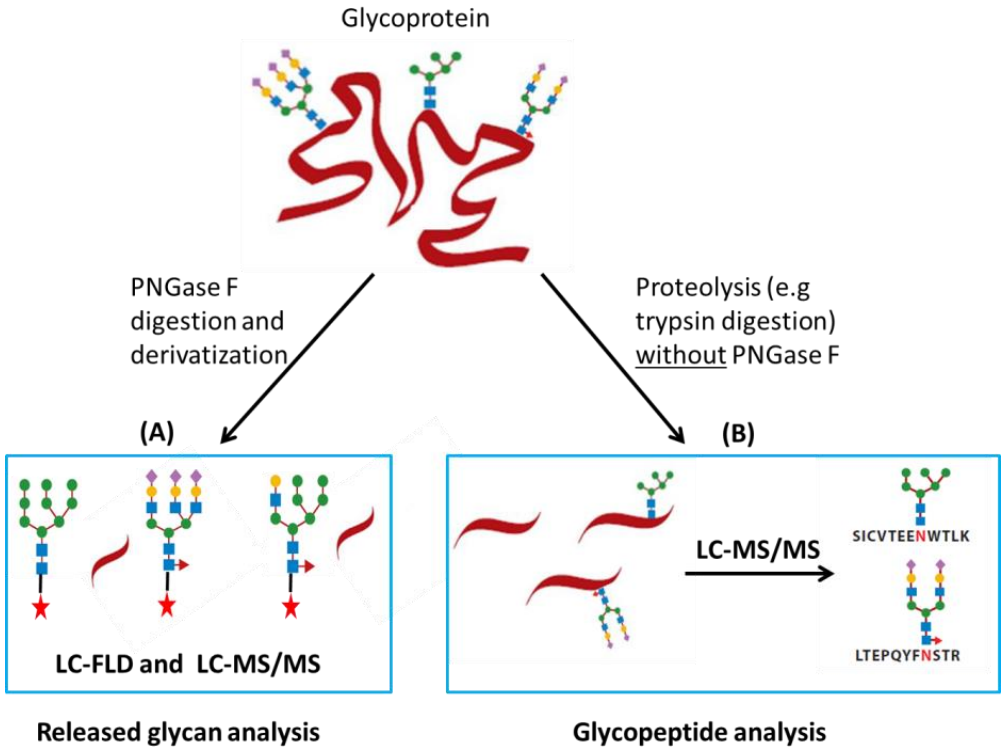


Figure 5. Representation of glycosylation analysis by released glycan approach **(A)** and glycopeptide approach **(B)**. Red stars indicate derivatization of released glycans. The key of the glycan compositions (monosaccharide residues) is shown in Figure 4.

Conversely, glycopeptide analysis (Figure 5B) provides site-specific glycosylation information, and it is the preferred method for glycosylation profiling of proteins containing more than one occupied glycosylation sites. The general approach for glycopeptide analysis involves enzymatic digestion of the glycoprotein (without glycan release) into peptides and glycopeptides, followed by electrospray ionization (ESI) or matrix assisted laser desorption ionization (MALDI) and mass spectrometry (MS) and/or tandem mass spectrometry (MS/MS) analyses.[133] Typical enzymes used for glycoprotein digestion include trypsin and Lys-C. Liquid chromatographic separation of the peptides and glycopeptides by reverse phase columns (e.g C_8

and C_{18} columns) prior to MS and MS/MS analysis is the most widely used method for glycopeptide profiling.

In cases where the glycoprotein of interest is in a biological matrix that contains other proteins and glycoproteins, purification or enrichment of the glycoprotein is a vital step prior to glycan or glycopeptide characterization. A number of methods, including lectin affinity, gel electrophoresis, antibody binding, and solid phase extraction can be used for glycoprotein purification.[138-141]

1.7.2.1 Glycosylation Analysis Using Mass Spectrometry

Mass spectrometry is a powerful analytical method for glycosylation characterization. Glycans and glycopeptides can be analyzed using MALDI-MS, LC-ESI-MS and MS/MS, or LC-FLD (for glycan analysis only).[142-144] However, LC-ESI-MS/MS is the preferred characterization method, especially for glycopeptide assignments, as separation and fragmentation of glycans and glycopeptides provides unambiguous assignments. For LC-FLD characterization of fluorescent-labeled glycans, mass spectrometry is typically used during method development to confirm the identities of the glycan peaks in the fluorescence profile (LC-FLD-MS).[135] Hence, once the glycan peak identities in the fluorescence profile of a particular protein have been confidently assigned using mass spectrometry, subsequent routine glycan analysis of the protein can be done using only LC-FLD without mass spectrometric detection.[136]

For glycopeptide analysis, high resolution MS and MS/MS data are typically used to determine the glycopeptide compositions, since MS data alone are not sufficient to assign isobaric glycopeptides. Glycopeptide analysis by tandem mass spectrometry can be done using a

variety of fragmentation techniques including electron transfer dissociation (ETD),[145, 146] collision induced dissociation (CID),[147] higher energy collision dissociation (HCD),[148] ultra violet photodissociatioon (UVPD),[149] and electron transfer higher energy collision induced dissociation (EThCD).[150] A combination of more than one fragmentation technique provides thorough glycopeptide characterization.[148, 151] Among these fragmentation techniques, ETD and CID are the most widely used dissociation methods.

Glycopeptide analysis using both ETD and CID is highly desirable for unambiguous glycopeptide assignments because the two fragmentation techniques give complementary glycopeptide information. Figure 6 shows the characteristic dissociation behaviors of glycopeptides when subjected to ETD and CID conditions. Under ETD fragmentation conditions, the peptide components of glycopeptide compositions are cleaved while the glycan components remain intact (Figure 6A).[146] Hence, ETD provides information about the peptide component of a glycopeptide composition without any information on the glycan component. On the contrary, glycosidic bonds are cleaved during CID fragmentation (Figure 6B) leading to dissociation of the glycans, and information about a glycopeptide's glycan component can be obtained.[147] In rare cases where unusually high CID energy is used, both the glycan and peptide components fragment, but the CID data are usually more complex and difficult to assign in such cases.[152, 153] Generally, the cleavage of a glycopeptide's glycosidic bonds during CID leads to loss of mono-, di-, tri-, and tetra- saccharide residues from the glycan component, giving rise to oxonium ions at *m/z* 366, 528, 690, and 657 (for glycopeptides containing sialic acids). In addition, CID spectra of glycopeptide compositions typically contain Y1 ions (peptide + GlcNAc ion)[154] that can be used to determine the mass of the peptide portion of the glycopeptide. The

presence of the oxonium and Y1 ions in CID data indicates that the precursor ion is a glycopeptide. Nonetheless, the interpretation of glycopeptide data into meaningful glycosylation information is still very challenging.

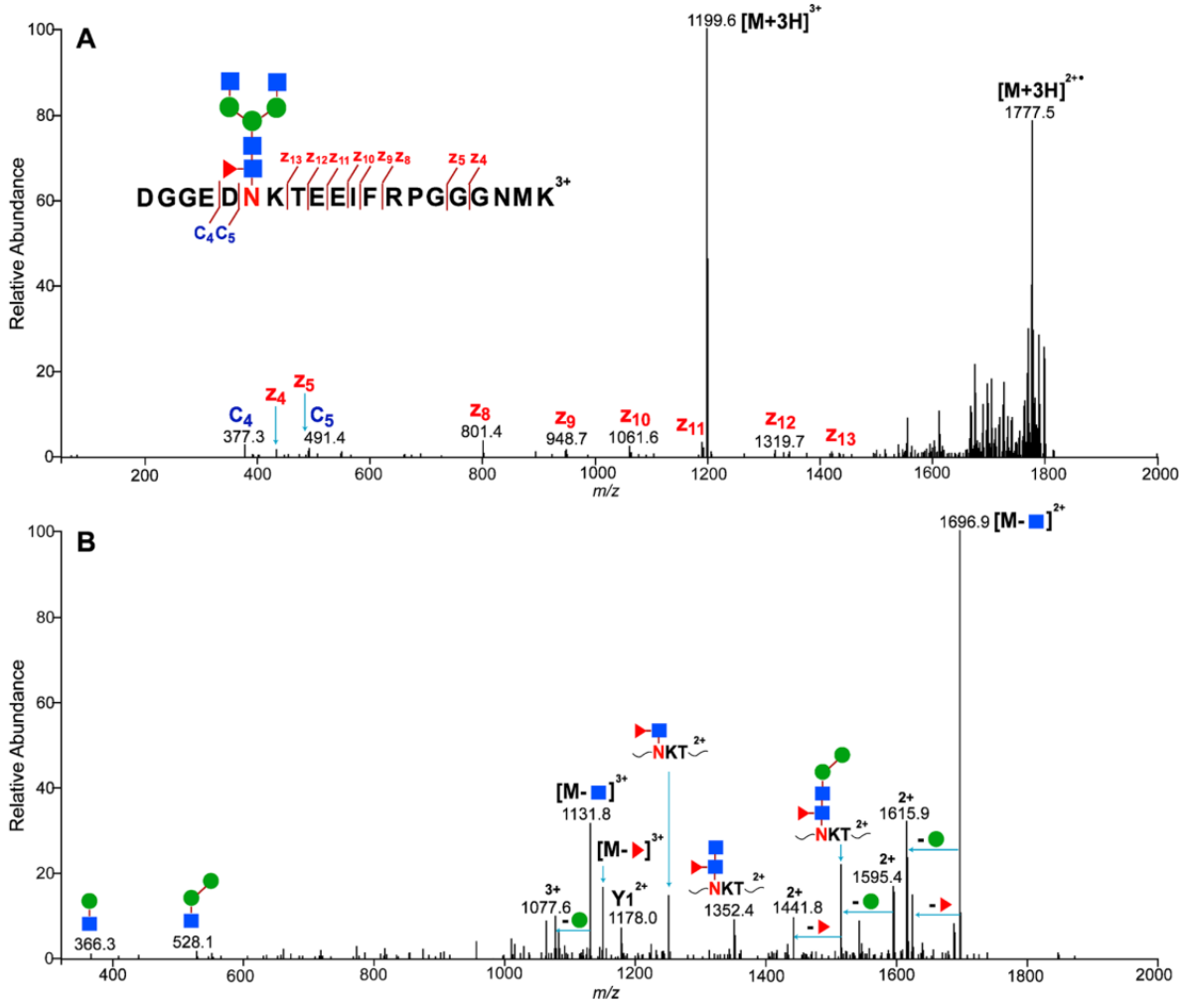


Figure 6. Representative spectra showing fragmentation of the same glycopeptide by ETD **(A)** and CID **(B)**. During ETD fragmentation, the glycan component remains intact while the peptide component is cleaved, and during CID the glycan component is cleaved while the peptide component is typically not cleaved. Oxonium ion peaks at *m/z* 366 and 528 are observed in the CID spectrum **(B)**. The blue squares, green circles, and red triangles denote N-acetylhexoseamine, hexose, and fucose monosaccharide residues, respectively. This figure is reproduced from Reference 145 with permission from the *American Chemical Society*.

1.7.3 Automated tools for Glycosylation Analysis

Manual glycosylation characterization is a time-consuming and difficult task that requires expertise in the glycomics or glycopeptide analysis fields. Manual analysis becomes even more difficult when large cohorts of glycoproteins need to be routinely analyzed, as is the case in biopharmaceutical industries. Hence, numerous automated tools have been developed for rapid and high-throughput glycosylation analysis. Besides decreasing analysis time, automated tools are also advantageous because end-users of the tools don't have to be experts in glycan or glycopeptide analysis in order to use them.

A number of glycomics tools that utilize MS and MS/MS data for glycosylation profiling in proteins have been reported in the literature. Goldberg *et al*[155] developed one of the early glycan annotation tools (Cartoonist) for automated assignment of released and permethylated *N*-glycans using MALDI-MS data. The tool provides plausible glycan compositions from a database of structures that match the masses of peaks in the MALDI-MS spectra. Glycoworkbench functions in a similar manner, but instead of matching experimental peaks to glycan structures in a database, the peaks are matched to plausible glycan compositions provided by the user.[156] Hu *et al* recently developed MultiGlycan-ESI,[157] a tool that assigns permethylated glycans using LC-ESI-MS data. Several automated glycomics algorithms for glycan assignment using MS/MS data have also been reported, and they have recently been reviewed by Tsai *et al*.[158]

For automated glycopeptide analysis, GlycoMod[159] and Glycopep DB[160] are online tools that can be used to obtain potential glycopeptide compositions of a protein using MS data. However, these tools usually suggest a number of isobaric glycopeptide compositions for each

MS ion. Hence, further analysis using algorithms that can assign glycopeptides from MS/MS data is required for confident glycopeptide assignment. Recently, numerous tools have been developed to meet this need. For example, SweetNET,[161] pGlyco,[162] MAGIC,[163] and GPQuest[164] are a fraction of software developed within the past three years (2015 to 2017) for glycopeptide analysis using MS/MS data. Generally, the tools assign glycopeptides to MS/MS data by comparing *in silico* spectra of the potential glycopeptide candidates to the experimental MS/MS spectra, and the potential glycopeptide candidate that best matches the experimental spectrum is assigned to the spectrum. A number of recent review articles that describe these tools and other automated glycopeptide software are available in the literature.[158, 165, 166]

The majority of the methods for glycopeptide analysis, including automated methods, are based on the well-established ETD and CID fragmentation techniques. Although these methods have significantly advanced the field of glycosylation characterization, development of high quality automated tools is still very challenging, and new approaches are needed to verify the accuracy of the automated tools in order to increase confidence in the results. Also, new automated methods are expected in the near future to accommodate fragmentation techniques like UVPD, HCD, and EThcD.

1.8 Summary of Subsequent Chapters

Chapter 2 describes a study to investigate the presence of disulfide bond-mediated structural isomers of endogenous IgG3 antibodies. A method combining ETD and CID data to map the disulfide bonds in proteins is used. Besides the classical IgG3 disulfide bonds, no alternative disulfide linkages were identified in the IgG3 molecules. This result indicates that unlike IgG2 and IgG4 that have isoforms resulting from alternative disulfide bond patterns,

native IgG3, like endogenous IgG1, does not have disulfide isoforms. In addition, the study demonstrates that when using ETD for disulfide bond analysis, an additional CID-MS^3 step is not always necessary to confirm the amino acid sequences of disulfide-linked peptides, since product ions (c/z ions) from the disulfide-bonded peptides can be readily identified in the ETD spectra.

Chapter 3 describes two new tools: *Glycopeptide Decoy Generator* and a large set of manually assigned glycopeptide CID data that ensure rapid assessment of existing scoring algorithms designed for glycopeptide identification from CID data. The tools also enable accurate development of CID glycopeptide software. *Glycopeptide Decoy Generator* generates abundant glycopeptide decoys *de novo* for any target glycopeptide composition. This work is important because developing high-quality scoring algorithms for the assignment of glycopeptides to MS/MS data is a challenging task and verifying the accuracy of the results from glycopeptide tools is vital in order to increase reliability on the tools. As a demonstration of the relevance of the new tools, they were used to assess the accuracy of GlycoPep Grader, an existing software that assigns glycopeptides to CID spectra. The results show that GlycoPep Grader has some limitations in scoring spectra from fucosylated glycopeptides, and these weaknesses were only identified when abundant decoys were generated by the *decoy generator* for each fucosylated target glycopeptide and scored against the CID spectra of the targets from the large dataset of glycopeptide spectra. These new tools are a useful contribution to the area of glycopeptide bioinformatics software development as they can advance the field by facilitating development of highly accurate scoring algorithms.

Chapter 4 describes a systematic study carried out to address GlycoPep Grader's limitations in scoring fucosylated glycopeptides, and outlines proposals for improving the tool. The approach that was used to identify the root cause of GlycoPep Grader's weaknesses can be applied to other glycopeptide scoring algorithm. After studying the CID fragmentation characteristics of fucosylated glycopeptides, product ions from an additional fragmentation pathway were identified and these ions were not incorporated into the original GlycoPep Grader scoring rules. Potential updates that take these new product ions into account are proposed.

1.9 References

1. Woycechowsky, K. J.; Raines, R. T., *Curr. Opin. Chem. Biol.*, **2000,** *4* (5), 533-539.
2. Wilkinson, B.; Gilbert, H. F., *Biochim. Biophys. Acta.*, **2004,** *1699* (1-2), 35-44.
3. Butera, D.; Cook, K. M.; Chiu, J.; Wong, J. W.; Hogg, P. J., *Blood,* **2014,** *123* (13), 2000-2007.
4. Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Campbell, D. S.; Sun, Z.; Bletz, J. A.; Mallick, P.; Katz, J. E.; Malmstrom, J.; Ossola, R.; Watts, J. D.; Lin, B.; Zhang, H.; Moritz, R. L.; Aebersold, R., *Mol. Cell. Proteomics,* **2011,** *10* (9), M110.006353.
5. Wypych, J.; Li, M.; Guo, A.; Zhang, Z.; Martinez, T.; Allen, M. J.; Fodor, S.; Kelner, D. N.; Flynn, G. C.; Liu, Y. D.; Bondarenko, P. V.; Ricci, M. S.; Dillon, T. M.; Balland, A., *J. Biol. Chem.*, **2008,** *283* (23), 16194-16205.
6. Alewood, D.; Nielsen, K.; Alewood, P. F.; Craik, D. J.; Andrews, P.; Nerrie, M.; White, S.; Domagala, T.; Walker, F.; Rothacker, J.; Burgess, A. W.; Nice, E. C., *Growth Factors,* **2005,** *23* (2), 97-110.
7. Zhang, L.; Chou, C. P.; Moo-Young, M., *Biotechnol. Adv.*, **2011,** *29* (6), 923-929.

8. Perry, L. J.; Wetzel, R., *Science,* **1984,** *226* (4674), 555-557.

9. Mansfeld, J.; Vriend, G.; Dijkstra, B. W.; Veltman, O. R.; Van den Burg, B.; Venema, G.; Ulbrich-Hofmann, R.; Eijsink, V. G., *J. Biol. Chem.,* **1997,** *272* (17), 11152-11156.

10. Kim, D. Y.; Kandalaft, H.; Ding, W.; Ryan, S.; van Faassen, H.; Hirama, T.; Foote, S. J.; MacKenzie, R.; Tanha, J., *Protein Eng. Des. Sel.,* **2012,** *25* (10), 581-590.

11. Dombkowski, A. A.; Sultana, K. Z.; Craig, D. B., *FEBS Lett.,* **2014,** *588* (2), 206-212.

12. Seger, S. T.; Breinholt, J.; Faber, J. H.; Andersen, M. D.; Wiberg, C.; Schjodt, C. B.; Rand, K. D., *Anal. Chem.,* **2015,** *87* (12), 5973-5980.

13. Go, E. P.; Cupo, A.; Ringe, R.; Pugach, P.; Moore, J. P.; Desaire, H., *J. Virol.,* **2015,** *90* (6), 2884-2894.

14. Le, Q. A. T.; Joo, J. C.; Yoo, Y. J.; Kim, Y. H., *Biotechnol. Bioeng.,* **2012,** *109* (4), 867-876.

15. Compton, J. R.; Legler, P. M.; Clingan, B. V.; Olson, M. A.; Millard, C. B., *Proteins,* **2011,** *79* (4), 1048-1060.

16. Liu, W.; Onda, M.; Kim, C.; Xiang, L.; Weldon, J. E.; Lee, B.; Pastan, I., *Protein Eng. Des. Sel.,* **2012,** *25* (1), 1-6.

17. Li, Y.; Li, X.; Zheng, X.; Tang, L.; Xu, W.; Gong, M., *Peptides,* **2011,** *32* (7), 1400-1407.

18. Li, Y.; Zheng, X.; Tang, L.; Xu, W.; Gong, M., *Peptides,* **2011,** *32* (6), 1303-1312.

19. Siadat, O. R.; Lougarre, A.; Lamouroux, L.; Ladurantie, C.; Fournier, D., *BMC Biochem.,* **2006,** *7*, 12.

20. Azimi, I.; Wong, J. W.; Hogg, P. J., *Antioxid.Redox Signal.,* **2011,** *14* (1), 113-126.

21. Hogg, P. J., *Nat. Rev. Cancer,* **2013,** *13* (6), 425-431.

22. Suzuki, M.; Yamanoi, A.; Machino, Y.; Kobayashi, E.; Fukuchi, K.; Tsukimoto, M.; Kojima, S.; Kohroki, J.; Akimoto, K.; Masuho, Y., *Biochem. Biophys. Res. Commun.,* **2013,** *436* (3), 519-524.

23. Suzuki, M.; Yamanoi, A.; Machino, Y.; Ootsubo, M.; Izawa, K.; Kohroki, J.; Masuho, Y., *J. Biochem.,* **2016,** *159* (1), 67-76.

24. Lightle, S.; Aykent, S.; Lacher, N.; Mitaksov, V.; Wells, K.; Zobel, J.; Oliphant, T., *Protein sci.,* **2010,** *19* (4), 753-762.

25. Walsh, G., *Nat. Biotechnol.,* **2014,** *32* (10), 992-1000.

26. Reichert, J. M., *mAbs.,* **2016,** *8* (2), 197-204.

27. Michel, M. L.; Tiollais, P., *Pathol. Biol.,* **2010,** *58* (4), 288-295.

28. Nascimento, I. P.; Leite, L. C. C., *Braz. J. Med. Biol. Res.,* **2012,** *45* (12), 1102-1111.

29. Amoresano, A.; Orru, S.; Siciliano, R. A.; De Luca, E.; Napoleoni, R.; Sirna, A.; Pucci, P., *Biol. Chem.,* **2001,** *382* (6), 961-968.

30. Reichert, J. M., *mAbs.,* **2012,** *4* (3), 413-415.

31. Wadhwa, M.; Thorpe, R., *Thromb. Haemost.,* **2008,** *99* (5), 863-873.

32. Calcutt, N. A.; Jolivalt, C. G.; Fernyhough, P., *Curr. Drug Targets,* **2008,** *9* (1), 47-59.

33. Liu, H. C.; May, K., *mAbs.,* **2012,** *4* (1), 17-23.

34. Trexler-Schmidt, M.; Sargis, S.; Chiu, J.; Sze-Khoo, S.; Mun, M.; Kao, Y. H.; Laird, M. W., *Biotechnol. Bioeng.,* **2010,** *106* (3), 452-461.

35. Hutterer, K. M.; Hong, R. W.; Lull, J.; Zhao, X.; Wang, T.; Pei, R.; Le, M. E.; Borisov, O.; Piper, R.; Liu, Y. D.; Petty, K.; Apostol, I.; Flynn, G. C., *mAbs.,* **2013,** *5* (4), 608-613.

36. Yoshioka, S.; Aso, Y.; Izutsu, K.; Terao, T., *Pharm. Res.,* **1993,** *10* (5), 687-691.

37. Chandrasekhar, S.; Topp, E. M., *J. Pharm. Sci.,* **2015,** *104* (4), 1291-1302.

38. Andya, J. D.; Hsu, C. C.; Shire, S. J., *AAPS PharmSci.,* **2003,** *5* (2), E10.

39. Vazquez-Rey, M.; Lang, D. A., *Biotechnol. Bioeng.,* **2011,** *108* (7), 1494-508.

40. FDA, Q6B Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological Products. **199**.

41. Rathore, A. S.; Winkle, H., *Nat. Biotech.,* **2009,** *27* (1), 26-34.

42. Klaus, W.; Broger, C.; Gerber, P.; Senn, H., *J. Mol. Biol.,* **1993,** *232* (3), 897-906.

43. Mobli, M.; King, G. F., *Toxicon.,* **2010,** *56* (6), 849-854.

44. Poppe, L.; Hui, J. O.; Ligutti, J.; Murray, J. K.; Schnier, P. D., *Anal. Chem.,* **2012,** *84* (1), 262-266.

45. Liu, D.; Cowburn, D., *Biophys. Rep.,* **2016,** *2* (1), 33-43.

46. Zhang, B.; Cockrill, S. L., *Anal. Chem.,* **2009,** *81* (17), 7314-7320.

47. Perham, R. N., *Biochem. J.,* **1967,** *105* (3), 1203-1207.

48. Milstein, C.; Frangion.B, *Biochem. J.,* **1971,** *121* (2), 217-225.

49. Gorman, J. J.; Wallis, T. P.; Pitt, J. J., *Mass Spectrom. Rev.,* **2002,** *21* (3), 183-216.

50. Tsai, P. L.; Chen, S. F.; Huang, S. Y., *Rev. Anal. Chem.,* **2013,** *32* (4), 257-268.

51. Wiesner, J.; Resemann, A.; Evans, C.; Suckau, D.; Jabs, W., *Expert Rev. Proteomics,* **2015,** *12* (2), 115-123.

52. Wu, S.-L.; Jiang, H.; Lu, Q.; Dai, S.; Hancock, W.; Karger, B. L., *Anal. Chem.,* **2009,** *81*, 112-122.

53. Clark, D. F.; Go, E. P.; Desaire, H., *Anal. Chem.,* **2013,** *85* (2), 1192-1199.

54. Lakbub, J. C.; Clark, D. F.; Shah, I. S.; Zhu, Z.; Su, X.; Go, E. P.; Tolbert, T. J.; Desaire, H., *Anal. Methods,* **2016,** *8* (31), 6046-6055.

55. Pitt, J. J.; Da Silva, E.; Gorman, J. J., *J. Biol. Chem.,* **2000,** *275* (9), 6469-6478.

56. Zhang, W.; Marzilli, L. A.; Rouse, J. C.; Czupryn, M. J., *Anal. Biochem.* **2002,** *311* (1), 1-9.

57. Sung, W. C.; Chang, C. W.; Huang, S. Y.; Wei, T. Y.; Huang, Y. L.; Lin, Y. H.; Chen, H. M.; Chen, S. F., *Biochim. Biophys. Acta* **2016,** *1864* (9), 1188-1194.

58. Wang, Y.; Li, H.; Shameem, M.; Xu, W., *Anal. Biochem.*, **2016,** *495*, 21-28.

59. Sanger, F., *Nature,* **1953,** *171* (4362), 1025-1026.

60. Ryle, A. P.; Sanger, F., *Biochem. J.,* **1955,** *60* (4), 535-540.

61. Costantino, H. R.; Schwendeman, S. P.; Langer, R.; Klibanov, A. M., *Biochem. (Mosc),* **1998,** *63* (3), 357-363.

62. Chandrasekhar, S.; Moorthy, B. S.; Xie, R.; Topp, E. M., *Pharm. Res.*, **2016,** *33* (6), 1370-1382.

63. Monahan, F. J.; German, J. B.; Kinsella, J. E., *J. Agr. Food Chem.,* **1995,** *43* (1), 46-52.

64. Kerr, J.; Schlosser, J. L.; Griffin, D. R.; Wong, D. Y.; Kasko, A. M., *Biomacromolecules,* **2013,** *14* (8), 2822-2829.

65. Wang, Y.; Lu, Q.; Wu, S. L.; Karger, B. L.; Hancock, W. S., *Anai. Chem.,* **2011,** *83* (8), 3133-3140.

66. Creamer, L. K.; Bienvenue, A.; Nilsson, H.; Paulsson, M.; van Wanroij, M.; Lowe, E. K.; Anema, S. G.; Boland, M. J.; Jimenez-Flores, R., *J. Agr. Food Chem.,* **2004,** *52* (25), 7660-7668.

67. Sechi, S.; Chait, B. T., *Anal. Chem.,* **1998,** *70* (24), 5150-5158.

68. Rogers, L. K.; Leinweber, B. L.; Smith, C. V., *Anal. Biochem.*, **2006,** *358* (2), 171-184.

69. Lu, S.; Fan, S.-B.; Yang, B.; Li, Y.-X.; Meng, J.-M.; Wu, L.; Li, P.; Zhang, K.; Zhang, M.-J.; Fu, Y.; Luo, J.; Sun, R.-X.; He, S.-M.; Dong, M.-Q., *Nat. Methods,* **2015,** *12* (4), 329-331.

70. Switzar, L.; Giera, M.; Niessen, W. M. A., *J. Proteome Res.*, **2013,** *12* (3), 1067-1077.

71. Ni, W.; Lin, M.; Salinas, P.; Savickas, P.; Wu, S. L.; Karger, B. L., *J. Am. Soc. Mass Spectrom.,* **2013,** *24* (1), 125-133.

72. Na, S.; Paek, E.; Choi, J.-S.; Kim, D.; Lee, S. J.; Kwon, J., *Mol. Biosyst.*, **2015,** *11* (4), 1156-1164.

73. Glatter, T.; Ludwig, C.; Ahrné, E.; Aebersold, R.; Heck, A. J. R.; Schmidt, A., *J. Proteome Res.*, **2012,** *11* (11), 5145-5156.

74. Liu, F.; van Breukelen, B.; Heck, A. J., *Mol. Cell. Proteomics,* **2014,** *13* (10), 2776-2786.

75. Go, E. P.; Hua, D.; Desaire, H., *J. Proteome Res.* **2014,** *13* (9), 4012-4027.

76. Moulaei, T.; Stuchlik, O.; Reed, M.; Yuan, W.; Pohl, J.; Lu, W.; Haugh-Krumpe, L.; O'Keefe, B. R.; Wlodawer, A., *Protein Sci.*, **2010,** *19* (9), 1649-1661.

77. Pike, G. M.; Madden, B. J.; Melder, D. C.; Charlesworth, M. C.; Federspiel, M. J., *J. Biol. Chem.*, **2011,** *286* (20), 17954-17967.

78. Eden P. Go , Y. Z., Sushma Menon , and Heather Desaire., *J. Proteome Res.,* **2011,** *10* (2), 578-591.

79. Cramer, C. N.; Kelstrup, C. D.; Olsen, J. V.; Haselmann, K. F.; Nielsen, P. K., *Anal. Chem.*, **2017,** *89* (11), 5949-5957.

80. Clark, D. F.; Go, E. P.; Toumi, M. L.; Desaire, H., *J. Am. Soc. Mass Spectrom.,* **2011,** *22* (3), 492-498.

81. Wu, S.-L.; Jiang, H.; Hancock, W.; Karger, B. L., *Anal. Chem.,* **2010,** *82,* 5296–5303.

82. Choi, S.; Jeong, J.; Na, S.; Lee, H. S.; Kim, H.-Y.; Lee, K.-J.; Paek, E., *J. Proteome Res.,* **2010,** *9* (1), 626-635.

83. Chrisman, P. A.; Pitteri, S. J.; Hogan, J. M.; McLuckey, S. A., *J. Am. Soc. Mass Spectrom.,* **2005,** *16* (7), 1020-1030.

84. Gunawardena, H. P.; Gorenstein, L.; Erickson, D. E.; Xia, Y.; McLuckey, S. A., *Int. J. Mass Spectrom.,* **2007,** *265* (2-3), 130-138.

85. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., *Nat. Methods,* **2007,** *4* (9), 709-712.

86. Frese, C. K.; Altelaar, A. F. M.; van den Toorn, H.; Nolting, D.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S., *Anal. Chem.,* **2012,** *84* (22), 9668-9673.

87. Brunner, A. M.; Lössl, P.; Liu, F.; Huguet, R.; Mullen, C.; Yamashita, M.; Zabrouskov, V.; Makarov, A.; Altelaar, A. F. M.; Heck, A. J. R., *Anal. Chem.,* **2015,** *87* (8), 4152-4158.

88. Frese, C. K.; Altelaar, A. F. M.; Hennrich, M. L.; Nolting, D.; Zeller, M.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S., *J. Proteome Res.,* **2011,** *10* (5), 2377-2388.

89. Bookwalter, C. W.; Zoller, D. L.; Ross, P. L.; Johnston, M. V., *J. Am. Soc. Mass Spectrom.,* **1995,** *6* (9), 872-876.

90. Agarwal, A.; Diedrich, J. K.; Julian, R. R., *Anal. Chem.,* **2011,** *83* (17), 6455-6458.

91. Cole, S. R.; Ma, X.; Zhang, X.; Xia, Y., *J. Am. Soc. Mass Spectrom.,* **2012,** *23* (2), 310-320.

92. Durand, K. L.; Tan, L.; Stinson, C. A.; Love-Nkansah, C. B.; Ma, X.; Xia, Y., *J. Am. Soc. Mass Spectrom.,* **2017**, 1-10.

93. Massonnet, P.; Upert, G.; Smargiasso, N.; Gilles, N.; Quinton, L.; De Pauw, E., *Anal. Chem.,* **2015,** *87* (10), 5240-5246.

94. Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E. P.; Shabanowitz, J.; Hunt, D. F., *Biochim. Biophys. Acta.,* **2006,** *1764* (12), 1811-1822.

95. Lu, C.; Liu, D.; Liu, H.; Motchnik, P., *mAbs.,* **2013,** *5* (1), 102-113.

96. Klapoetke, S.; Xie, M. H., *J. Pharmaceut. Biomed.,* **2017,** *132*, 238-246.

97. Li, X.; Xu, W.; Paporello, B.; Richardson, D.; Liu, H., *Anal. Biochem.,* **2013,** *439* (2), 184-186.

98. Li, X.; Wang, F.; Xu, W.; May, K.; Richardson, D.; Liu, H., *Anal. Biochem.,* **2013,** *436* (2), 93-100.

99. Liu, H.; Lei, Q. P.; Washabaugh, M., *Anal. Chem.,* **2016,** *88* (10), 5080-5087.

100. Cramer, C. N.; Haselmann, K. F.; Olsen, J. V.; Nielsen, P. K., *Anal. Chem.,* **2016,** *88* (3), 1585-1592.

101. Switzar, L.; Nicolardi, S.; Rutten, J. W.; Oberstein, S. A.; Aartsma-Rus, A.; van der Burgt, Y. E., *J. Am. Soc. Mass Spectrom.,* **2016,** *27* (1), 50-58.

102. Zheng, Q.; Zhang, H.; Chen, H., *Int. J. Mass Spectrom.,* **2013,** *353*, 84-92.

103. Zhang, Y.; Dewald, H. D.; Chen, H., *J. Proteome Res.,* **2011,** *10* (3), 1293-1304.

104. Jiang, H.; Wu, S.-L.; Karger, B. L.; Hancock, W. S., *Biotechnol. Prog.,* **2009,** *25* (1), 207-218.

105. Xiang, T.; Chumsae, C.; Liu, H., *Anal. Chem.,* **2009,** *81*, 8101–8108

106. Chumsae, C.; Gaza-Bulseco, G.; Liu, H., *Anal. Chem.,* **2009,** *81*, 6449–6457.

107. Xu, H.; Zhang, L.; Freitas, M. A., *J. Proteome Res.,* **2008,** *7* (1), 138-144.

108. Murad, W.; Singh, R.; Yen, T. Y., *BMC Bioinformatics,* **2011,** *12 Suppl 1*, S12.

109. Murad, W.; Singh, R., *IEEE Trans. Nanobioscience,* **2013,** *12* (2), 69-71.

110. Huang, S. Y.; Wen, C. H.; Li, D. T.; Hsu, J. L.; Chen, C.; Shi, F. K.; Lin, Y. Y., *Anal. Chem.,* **2008,** *80* (23), 9135-9140.

111. Huang, S.-Y.; Hsieh, Y.-T.; Chen, C.-H.; Chen, C.-C.; Sung, W.-C.; Chou, M.-Y.; Chen, S.-F., *Anal. Chem.,* **2012,** *84* (11), 4900-4906.

112. Huang, S. Y.; Chen, S. F.; Chen, C. H.; Huang, H. W.; Wu, W. G.; Sung, W. C., *Anal. Chem.,* **2014,** *86* (17), 8742-8750.

113. Goyder, M. S.; Rebeaud, F.; Pfeifer, M. E.; Kalman, F., *Expert Rev. Proteomics,* **2013,** *10* (5), 489-501.

114. Bagal, D.; Valliere-Douglass, J. F.; Balland, A.; Schnier, P. D., *Anal. Chem.,* **2010,** *82* (16), 6751-6755.

115. Wang, Q.; Lacher, N. A.; Muralidhara, B. K.; Schlittler, M. R.; Aykent, S.; Demarest, C. W., *J. Sep. Sci.,* **2010,** *33* (17-18), 2671-2680.

116. Cao, X.; He, Y.; Smith, J.; Wirth, M. J., *J. Chromatogr. A,* **2015,** *1410*, 147-153.

117. Lodish, H.; Berk, A.; Zipursky, S. L.; Matsudaira, D.; Darnell, J., *Molecular Cell Biology*, 4th ed.; W. H. Freeman and Company, New York, **2000**, pp. 712-722.

118. Apweiler, R.; Hermjakob, H.; Sharon, N., *Biochim. Biophys. Acta,* **1999,** 1473 (1), 4-8.

119. Kowarik, M.; Numao, S.; Feldman, M. F.; Schulz, B. L.; Callewaert, N.; Kiermaier, E.; Catrein, I.; Aebi, M., *Science,* **2006,** 314 (5802), 1148-1150.

120. Van den Steen, P.; Rudd, P. M.; Dwek, R. A.; Opdenakker, G., *Crit. Rev. Biochem. Mol. Biol.,* **1998,** 33 (3), 151-208.

121. Tian, E.; Ten Hagen, K. G., *Glycoconj. J.*, **2009,** 26 (3), 325-334.

122. Hofsteenge, J.; Mueller, D. R.; de Beer, T.; Loeffler, A.; Richter, W. J.; Vliegenthart, J. F. G., *Biochemistry*, **1994,** 33 (46), 13524-13530.

123. Stepper, J.; Shastri, S.; Loo, T. S.; Preston, J. C.; Novak, P.; Man, P.; Moore, C. H.; Havlicek, V.; Patchett, M. L.; Norris, G. E., *FEBS Lett.*, **2011,** 585 (4), 645-50.

124. Gallagher, P.; Henneberry, J.; Wilson, I.; Sambrook, J.; Gething, M, J., *J. Cell Biol.*, **1988,** 107 (6), 2059-2073.

125. Helenius, A.; Aebi, M., *Science*, **2001,** 291 (5512), 2364-2369.

126. SolÁ, R. J.; Griebenow, K. A. I., *J. Pharm. Sci.*, **2009,** 98 (4), 1223-1245.

127. Li, Y.; Tian, Y.; Rezai, T.; Prakash, A.; Lopez, M. F.; Chan, D. W.; Zhang, H., *Anal. Chem.*, **2011,** 83 (1), 240-245.

128. Reichert, J. M., *mAbs*, **2016,** 8 (2), 197-204.

129. Xie, H.; Chakraborty, A.; Ahn, J.; Yu, Y. Q.; Dakshinamoorthy, D. P.; Gilar, M.; Chen, W.; Skilton, S. J.; Mazzeo, J. R., *mAbs* **2010,** 2 (4), 379-394.

130. Ivarsson, M.; Villiger, T. K.; Morbidelli, M.; Soos, M., *J. Biotechnol.*, **2014,** 188, 88-96.

131. van Berkel, P. H.; Gerritsen, J.; Perdok, G.; Valbjorn, J.; Vink, T.; van de Winkel, J. G.; Parren, P. W., *Biotechnol. Prog.*, **2009,** 25 (1), 244-251.

132. Planinc, A.; Dejaegher, B.; Vander Heyden, Y.; Viaene, J.; Van Praet, S.; Rappez, F.; Van Antwerpen, P.; Delporte, C., *Eur. J. Hosp. Pharm.*, **2016,** doi:10.1136/ejhpharm-2016-001022.

133. Zhu, Z.; Desaire, H., *Annu. Rev. Anal. Chem.*, **2015,** 8, 463-483.

134. Jensen, P. H.; Kolarich, D.; Packer, N. H., *FEBS J.*, **2010,** 277 (1), 81-94.

135. Lauber, M. A.; Yu, Y.-Q.; Brousmiche, D. W.; Hua, Z.; Koza, S. M.; Magnelli, P.; Guthrie, E.; Taron, C. H.; Fountain, K. J., *Anal. Chem.* **2015,** 87 (10), 5401-5409.

136. Aich, U.; Liu, A.; Lakbub, J.; Mozdzanowski, J.; Byrne, M.; Shah, N.; Galosy, S.; Patel, P.; Bam, N., *J. Pharm. Sci.,* **2016,** 105 (3), 1221-1232.

137. Burnina, I.; Hoyt, E.; Lynaugh, H.; Li, H.; Gong, B., *J. Chromatogr. A,* **2013,** 1307, 201-206.

138. Alley, W. R., Jr.; Mann, B. F.; Novotny, M. V., *Chem. Rev.,* **2013,** 113 (4), 2668-2732.

139. Hamada, H.; Tsuruo, T., *J. Biol. Chem.,* **1988,** 263 (3), 1454-1458.

140. Campbell, K. P.; MacLennan, D. H., *J. Biol. Chem.,* **1981,** 256 (9), 4626-4632.

141. Battle, K. N.; Jackson, J. M.; Witek, M. A.; Hupert, M. L.; Hunsucker, S. A.; Armistead, P. M.; Soper, S. A., *Analyst,* **2014,** 139 (6), 1355-1363.

142. Thaysen-Andersen, M.; Mysling, S.; Hojrup, P., *Anal. Chem.,* **2009,** 81 (10), 3933-3943.

143. Nwosu, C. C.; Seipert, R. R.; Strum, J. S.; Hua, S. S.; An, H. J.; Zivkovic, A. M.; German, B. J.; Lebrilla, C. B., *J. Proteome Res.,* **2011,** 10 (5), 2612-2624.

144. Aich, U.; Lakbub, J.; Liu, A., *Electrophoresis,* **2016,** 37 (11), 1468-1488.

145. Zhu, Z.; Su, X.; Go, E. P.; Desaire, H., *Anal. Chem.,* **2014,** 86 (18), 9212-9219.

146. Zhu, Z.; Hua, D.; Clark, D. F.; Go, E. P.; Desaire, H., *Anal. Chem.,* **2013,** 85 (10), 5023-5032.

147. Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H., *Anal. Chem.,* **2012,** 84 (11), 4821-4829.

148. Ford, K. L.; Zeng, W.; Heazlewood, J. L.; Bacic, A., *Front. Plant Sci.,* **2015,** 6, 674.

149. Madsen, J. A.; Ko, B. J.; Xu, H.; Iwashkiw, J. A.; Robotham, S. A.; Shaw, J. B.; Feldman, M. F.; Brodbelt, J. S., *Anal. Chem.,* **2013,** 85 (19), 9253-9261.

150. Yu, Q.; Wang, B.; Chen, Z.; Urabe, G.; Glover, M. S.; Shi, X.; Guo, L. W.; Kent, K. C.; Li, L., *J. Am. Soc. Mass Spectrom.* **2017**, doi:10.1007/s13361-017-1701-4

151. Go, E. P.; Herschhorn, A.; Gu, C.; Castillo-Menendez, L.; Zhang, S.; Mao, Y.; Chen, H.; Ding, H.; Wakefield, J. K.; Hua, D.; Liao, H. X.; Kappes, J. C.; Sodroski, J.; Desaire, H., *J. Virol.* **2015,** 89 (16), 8245-8257.

152. Vékey, K.; Ozohanics, O.; Tóth, E.; Jekő, A.; Révész, Á.; Krenyácz, J.; Drahos, L., *Int. J. Mass Spectrom.,* **2013,** 345–347, 71-79.

153. Kolli, V.; Dodds, E. D., *Analyst,* **2014,** 139 (9), 2144-2153.

154. Jiang, H.; Desaire, H.; Butnev, V. Y.; Bousfield, G. R., *J. Am. Soc. Mass Spectrom* **2004,** 15 (5), 750-758.

155. Goldberg, D.; Sutton-Smith, M.; Paulson, J.; Dell, A., *Proteomics,* **2005,** 5 (4), 865-875.

156. Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M., *J. Proteome Res.,* **2008,** 7 (4), 1650-1659.

157. Hu, Y.; Zhou, S.; Yu, C. Y.; Tang, H.; Mechref, Y., *Rapid comm. Mass Spectrom.,* **2015,** 29 (1), 135-142.

158. Tsai, P. L.; Chen, S. F., *Mass Spectrom. (Tokyo),* **2017,** 6 (Spec Iss), S0064.

159. Cooper, C. A.; Gasteiger, E.; Packer, N. H., *Proteomics,* **2001,** 1 (2), 340-349.

160. Go, E. P.; Rebecchi, K. R.; Dalpathado, D. S.; Bandu, M. L.; Zhang, Y.; Desaire, H., *Anal. Chem.,* **2007,** 79 (4), 1708-1713.

161. Nasir, W.; Toledo, A. G.; Noborn, F.; Nilsson, J.; Wang, M.; Bandeira, N.; Larson, G., *J. Proteome Res.,* **2016,** 15 (8), 2826-2840.

162. Zeng, W.-F.; Liu, M.-Q.; Zhang, Y.; Wu, J.-Q.; Fang, P.; Peng, C.; Nie, A.; Yan, G.; Cao, W.; Liu, C.; Chi, H.; Sun, R.-X.; Wong, C. C. L.; He, S.-M.; Yang, P., *Sci. Rep.,* **2016,** 6, 25102.

163. Lynn, K.-S.; Chen, C.-C.; Lih, T. M.; Cheng, C.-W.; Su, W.-C.; Chang, C.-H.; Cheng, C.-Y.; Hsu, W.-L.; Chen, Y.-J.; Sung, T.-Y., *Anal. Chem.,* **2015,** 87 (4), 2466-2473.

164. Toghi Eshghi, S.; Shah, P.; Yang, W.; Li, X.; Zhang, H., *Anal. Chem.,* **2015,** 87 (10), 5181-5188.

165. Woodin, C. L.; Maxon, M.; Desaire, H., *Analyst,* **2013,** 138 (10), 2793-2803.

166. Hu, H.; Khatri, K.; Zaia, J., *Mass Spectrom. Rev.,* **2017,** 36 (4), 475-498.

Chapter 2: Disulfide Bond Characterization of Endogenous IgG3 Monoclonal Antibodies Using LC-MS: An Investigation of IgG3 Disulfide-mediated Isoforms

*The work described in this chapter has been published in Anlytical Methods: Jude C. Lakbub, Daniel F. Clark, Ishan S. Shah, Zhikai Zhu, Xiaomeng Su, Eden P. Go, Thomas J. Tolbert and Heather Desaire. Anal. Methods, **2016**, 8, 6046-6055. It is reproduced here with permission from The Royal Society of Chemistry.*

Abstract

The use of monoclonal antibodies (mAbs) for the manufacture of innovator and biosimilar biotherapeutics has increased tremendously in recent years. From a structural perspective, mAbs have high disulfide bond content, and the correct disulfide connectivity is required for proper folding and to maintain their biological activity. Therefore, disulfide linkage mapping is an important component of mAb characterization for ensuring drug safety and efficacy. The native disulfide linkage patterns of all four subclasses of IgG antibodies have been well established since the late 1960s. Among these IgG subtypes, disulfide mediated isoforms have been identified for IgG2 and IgG4, and to a lesser extent in IgG1, which is the most studied IgG subclass. However, no studies have been carried out so far to investigate whether different IgG3 isoforms exist due to alternative disulfide connectivity. In an effort to investigate the presence of disulfide-mediated isoforms in IgG3, we employed a bottom-up mass spectrometry approach to accurately determine the disulfide bond linkages in endogenous human IgG3 monoclonal antibody. Our results show that no such alternative disulfide bonds exist. While

many antibody-based drugs are developed around IgG1, IgG3 represents a new, and in some cases, more desirable drug candidate. Our data represent the first demonstration that alternative disulfide bond arrangements are not present in endogenous IgG3; and therefore, they should not be present in recombinant forms used as antibody-based therapeutics.

2.1 Introduction

Human IgG3 monoclonal antibody is the most efficient IgG subclass in mediating effector functions, followed by IgG1, IgG2, and IgG4, respectively.[1,2] IgG3 displays the highest complement dependent cytotoxicity (CDC) and comparable antibody dependent cell-mediated cytotoxicity (ADCC) to IgG1, making it an ideal antibody drug candidate. Despite these ideal drug qualities and the rapidly growing use of monoclonal antibodies as biotherapeutics against various diseases, IgG3 is the only IgG subclass that has not yet been used for the production of antibody-based drugs. This is mainly due to its short half-life of seven days, compared to 21 days for the other IgG's.[3] This short half-life of IgG3 is generally attributed to its long hinge region of 62 amino acid residues, compared to 12 and 15 residues for the other IgG subclasses, making IgG3 more susceptible to proteolysis.[4] However, a recent report by Stapleton et al. showed that the short half-life is primarily due to the presence of arginine at position 435 (R435) of the IgG3 heavy chain, as opposed to histidine (H435) for the other IgG subclasses.[5] Mutation of the IgG3 heavy chain arginine 435 to histidine (H435–IgG3) extends the IgG3 half-life. This finding by Stapleton *et al* has energized interest in the production of IgG3-based biotherapeutics.

Consequently, timely studies on the structure and properties of endogenous IgG3 are now an urgent priority. For example, Plomp *et al* recently identified three O-glycosylation sites (each

having about 10% site occupancy) in the hinge region of endogenous IgG3 samples from six donors.[6] In a complementary line of work, we take on the challenge of investigating whether IgG3 is similar to the other IgG's in displaying endogenous isoforms resulting from alternative disulfide connectivity.

Disulfide bonds are vital post-translational modifications in therapeutic proteins, as they play a key role in mediating protein folding, stability and biological function.[7–9] The disulfide bond patterns of the four IgG subclasses (IgG1, IgG2, IgG3, and IgG4) were established in the late 1960s and early 1970s by Milstein et al. using diagonal paper electrophoresis and Edman degradation.[10–14] In addition to these classical IgG disulfide connections, alternative (non-classical) disulfide bonds have been identified in the constant regions of some IgG subclasses, leading to their disulfide-mediated structural isoforms. For example, in addition to the classical IgG4 structure with interchain disulfide bonds in the hinge region, IgG4 also forms intra-chain disulfide bonds in the hinge region, thereby forming an isoform that consists of two half molecules.[15,16] Additionally, both native and recombinant IgG2 antibodies have been shown to have two disulfide-mediated isoforms in addition to the classical IgG2 structure.[17–19] Furthermore, one report observed a trace amount of alternative intra-chain IgG1 disulfide bonds in the hinge region in addition to its conventional inter-chain disulfide bonds.[16] However, there is currently no study determining whether IgG3 also contains disulfide-mediated isoforms. This lack of information is likely because IgG3 has been overlooked as a promising drug candidate, due to its short half-life, which is also reflected by the lack of IgG3-based drugs in the market. With the discovery of H435–IgG3, which has comparable half-life to IgG1, IgG2, and IgG4, it is important to confirm the classical IgG3 disulfide bond connectivity and to determine whether or

not disulfide-mediated isoforms exist in endogenous IgG3. Data from such a study would facilitate future drug development work based on the IgG3 scaffold, because it would provide a blueprint for the appropriate disulfide bonding profile for recombinant IgG3-based therapeutics.

Herein, we use liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) to experimentally characterize the disulfide bond connectivity of IgG3 for the first time. Two batches of native (endogenous) IgG3 were obtained from two different sources and extensive disulfide bond characterization was done by expanding upon our previously published extracted ion chromatogram/Electron Transfer Dissociation (XIC/ETD) approach for rapid disulfide bond analysis in proteins.[20] With the combination of both the ETD-basedmethod and further analysis of collision induced dissociation (CID) data, all of the disulfide bonds in the constant region of the protein were accounted for. Using these techniques, we confirmed the classical disulfide bond pattern in the constant region of endogenous IgG3 antibodies and showed that unlike IgG2 and IgG4, which have well established conformational isoforms due to alternative disulfide bonds in their constant regions, endogenous IgG3 does not have any alternative disulfide bonds in its constant region and therefore does not have disulfide-mediated isoforms.

2.2 Experimental

2.2.1 Materials and Reagents

Tris(hydroxylmethyl)aminomethane (Trizma) base, guanidine hydrochloride, sodium acetate anhydrous, acetonitrile, N-ethylmaleimide, and gamma-globulins from human serum were purchased from Sigma-Aldrich (St. Louis, MO). High capacity protein A resin, sodium phosphate, glycine hydrochloride, sodium chloride, calcium chloride dihydrate, and optima grade

formic acid were from Fisher Scientific (Pittsburgh, PA). Native human IgG3 was purchased from Fitzgerald (Acton, MA), mouse anti-human IgG3 antibody was purchased from Invitrogen (ThermoFisher Scientific, Grand Island, NY), sequencing grade trypsin was acquired from Promega (Madison, WI), and protein G resin was prepared in-house.

2.2.2 Isolation of IgG3 from Human Gamma-globulins

Gamma-globulins from human serum were used as a source of serum IgG. IgG3 was isolated by sequential affinity purification using protein A and protein G. Protein A binds to IgG1, IgG2 and IgG4, and protein G binds to all IgG subclasses. About 100 mg of gamma globulins was re-suspended in 20 mM sodium phosphate, 150 mM sodium chloride, pH 7.0 (equilibration buffer) and passed over a protein A affinity column (5 mL bed volume of Pierce High-Capacity Protein A Ultralink resin) to capture all IgG subclasses except IgG3. The flow-through from protein A column (containing IgG3) was collected and the IgGs retained in the column were eluted with 0.1 M glycine hydrochloride, pH 2.5 (elution buffer). The column was re-equilibrated with phosphate buffer and the flow-through was reloaded onto the column. The procedure was repeated to remove any residual IgG subclasses other than IgG3. To further purify the IgG3, the final flow-through from the protein A column was loaded onto protein G column (5 mL bed volume of protein G resin), washed thoroughly with equilibration buffer (20x column volume), eluted with elution buffer, and immediately neutralized. IgG3 was dialyzed against the equilibration buffer, concentrated using Amicon ultra-centrifugal filters (30 kDa cutoff), and stored at -20 $^{\circ}$C. The isolated IgG3 was checked by SDS-PAGE and validated by Western blot using a mouse anti-human IgG3 antibody.

2.2.3 Proteolysis

To prevent disulfide bond shuffling, IgG3 samples were prepared under non-reducing conditions following a protocol modified from Reference 17. About 100 µg of IgG3 at a concentration of 1 µg/µL in 20 mM phosphate buffer (pH 7.0) was buffer exchanged using a 10 kDa molecular weight cut-off filter (Millipore, Billerica, MA) into 100 mM acetate buffer (pH 6.5) containing 7 M guanidine hydrochloride and 10 mM N-ethylmaleimide (NEM). The sample was incubated at 37 °C for two hours to allow for denaturation and capping of any free cysteine residues. After denaturation and alkylation, excess NEM and guanidine hydrochloride were removed by subjecting the samples to centrifugal filtration using a 10 kDa molecular weight cut-off filter, and the protein was reconstituted to a final concentration of 0.8 µg/µL in 100 mM Tris buffer (pH 7.0) containing 1 mM calcium chloride. Trypsin was added at an enzyme-to-protein ratio of 1:10 (w/w) and incubated for 15 hours at 37 °C. Tryptic digestion was stopped by adding 1 mL of formic acid for every 100 mL of solution. The digested IgG3 samples were diluted with water to a final concentration of 0.6 µg/µL and aliquots were stored at -20 °C until analysis. For the purpose of reproducibility, samples from the same IgG3 source (Sigma-Aldrich or Fitzgerald) were digested on two different days and each digested sample was run at least two times on different days using the same experimental procedure as described in the LC-MS analysis section.

2.2.4 LC-MS Analysis

Digested IgG3 samples were analyzed using reversed phase HPLC (Waters Acquity, Milford, MA) coupled with a LTQ Orbitrap Velos Pro hybrid mass spectrometer equipped with

ETD (Thermo Scientific, San Jose, CA). A solution of 5 mL of the tryptic digest was injected onto a C18 Aquasil Gold column (100 x 1 mm i.d, 175 A°, Thermo Scientific, San Jose, CA). The mobile phase A was 99.9% water with 0.1% formic acid; and mobile phase B was 99.9% acetonitrile with 0.1% formic acid. After sample injection, the tryptic peptides were eluted from the column at a flow rate of 50 µL/min using the following gradient: mobile phase B, initially held at 2% for 5 min, was increased to 35% in 55 min, and then ramped to 60% in 15 min, followed by a 10 minute isocratic elution at 95% B and re-equilibration.

Data acquisition was done in the data-dependent scan mode. After a survey MS scan from *m/z* 400 to 2000 in the Orbitrap mass analyzer at a resolution of 30 000 at *m/z* 400, the top 5 ions were sequentially selected for ETD (or CID, during CID experiments) in the linear ion trap. ETD and CID experiments were performed separately in different runs. For ETD experiments, charge state dependent ETD time and supplemental activation were enabled in order to enhance ETD efficiency. The ion–ion reaction time was maintained at 100 ms. For CID experiments, the activation time was set at 10 ms, and the normalization collision energy was 35%. The dynamic exclusion window and isolation width were set at 2 min and 2 Da, respectively, for both CID and ETD experiments. All data were collected in the positive ion mode with the ESI source spray voltage of 3.0 kV and capillary temperature of 250 °C. The data were acquired and analyzed using Xcalibur 2.7 software (ThermoElectron Corp, San Jose, CA).

2.3 Results and Discussion

2.3.1 Disulfide Analysis Approach

Most disulfide assignments were done manually using an augmented version of a method reported elsewhere.[20] A schematic representation of the disulfide mapping approach used to verify expected (classical) and alternative IgG3 disulfide bonds is shown in Figure 1A and B, respectively. To verify the classical disulfide linkage pattern, extracted ion chromatograms (XIC's) are constructed from ETD data based on the *m/z* values of two Cys-containing peptides that are expected to be linked through a disulfide bond (e.g. peptides P1 and P2, shown in Figure 1A). The XIC's of the two peptides are then compared to quickly verify whether the expected disulfide bond is present. If peaks having the same retention time (RT) are identified (such as the highlighted peaks in the figure) the peptides are preliminarily assigned to be disulfide-bonded partners, and the corresponding ETD spectrum is inspected. ETD preferentially cleaves disulfide bonds and produces intense peaks for each bonded peptide,[20,21] so the ETD spectrum is interrogated to determine whether intense marker ion peaks for peptides P1 and P2 are present, along with *c* and *z* fragment ions from each bonded chain and the intact disulfide-bonded peptide. If all these ions are present, the identity of the disulfide-linked peptides is assigned. Additionally, the assignment is quickly validated by matching the precursor ion mass to the theoretical mass of the dipeptide (the sum of the masses of the two Cys-containing peptides minus 2 Da).

After confirming the presence of the classical disulfide-linked peptides, alternatively connected disulfide-linked peptides are also searched for in a similar manner (Figure 1B). For example, in addition to the correct disulfide bond between P1 and P2, if peptide P1 is also

alternatively linked to peptide P3, the XIC's of both peptides would have peaks where the P1–P3 dipeptide (alternative disulfide bonded dipeptide) eluted. Therefore, by comparing the XIC's for peptides P1 and P3, and following the procedure described in the previous paragraph for identifying disulfide bonds between two peptides, the alternative disulfide bond between peptide P1 and peptide P3 could be identified, if it were present. Hence, to rapidly search for alternatively disulfide linked peptides, the XIC's of all Cys-containing peptides were compared, asking the question: are there any peaks that show up at the same retention time in at least two chromatograms, which could be aberrant disulfide linked peptides? When such peaks are identifiable, their ETD spectra are interrogated, as described above, to determine if the ion responsible for the peak is an aberrant disulfide-bonded peptide.

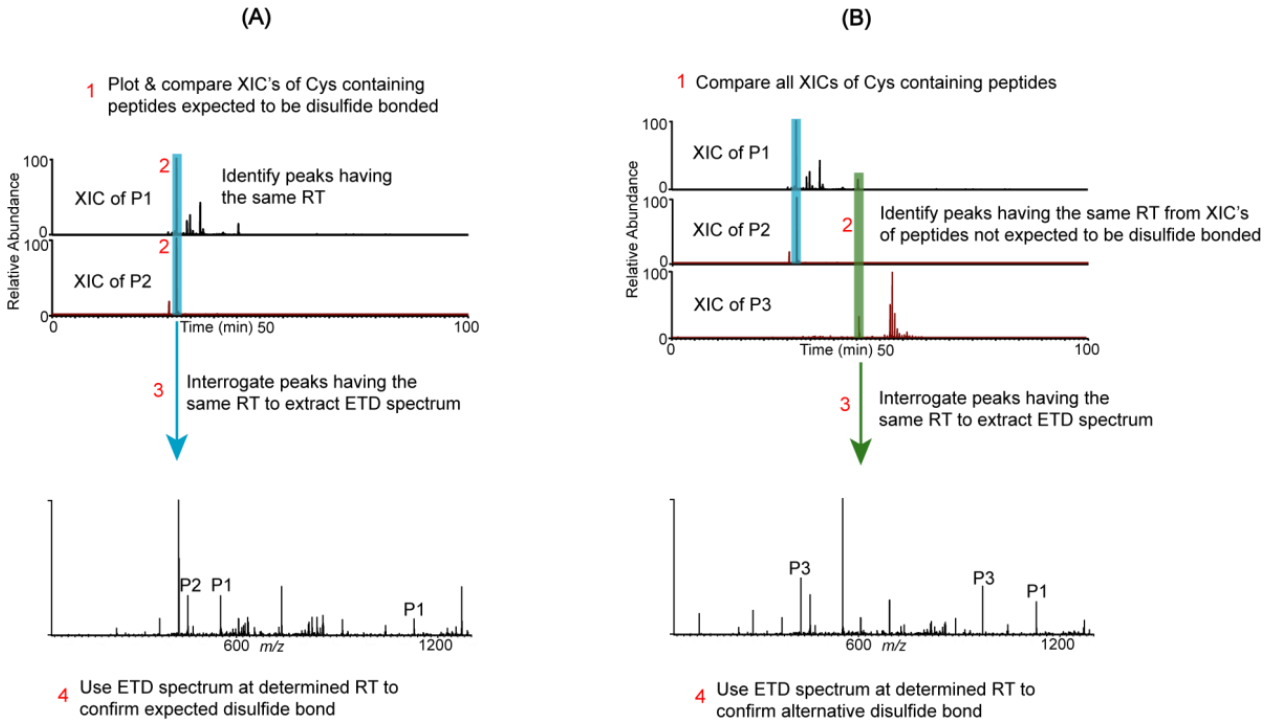


Figure 1. Schematic representation of the disulfide mapping approach for expected and alternative disulfides. **(A)** Assignment of an expected disulfide bond between Cys-containing peptides P1 and P2. Step 1: plot XIC's for each peptide. Step 2: identify peaks with the same retention time. Step 3: extract the corresponding ETD spectrum, which confirms the disulfide bond. Step 4: verify that marker ions of each chain (P1 & P2 peaks) and c and z ions from both chains are present in the ETD spectrum. **(B)** Alternative disulfide bonds are verified by the following: Step 1: plot and compare the XIC's of all Cys-containing peptides. Step 2: if peaks with the same RT are identified in the XIC's of peptides that are not expected to be disulfide bonded, steps 3 and 4 in **(A)** are used to verify whether the two peptides are bonded by an alternative disulfide bond.

2.3.2 Assignment of Expected (classical) IgG3 Disulfide Bonds

The classical disulfide bond structure of IgG3 is shown in Figure 2A. The comprehensive structure consists of 12 domains, two heavy chains and two light chains, with each chain having a variable and constant region. There are a total of 50 Cys residues that form 25 disulfide bonds: 21 in the constant regions and four in the variable regions. Because the IgG3 samples used in this study were isolated from human serum, the amino acid sequences of the variable regions were unknown; therefore, only the disulfide bonds in the constant regions were mapped. Additionally,

disulfide bonds for both the lambda and kappa light chains were mapped. The expected tryptic disulfide bonded peptides from the constant region are shown in Figure 2B.

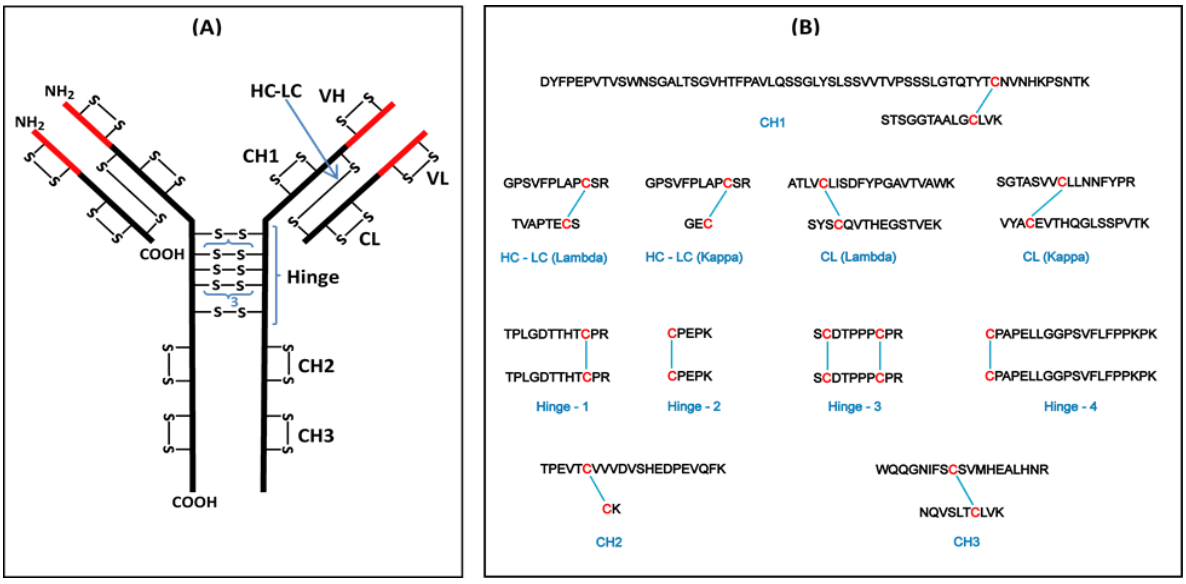


Figure 2. (A) Structure of a typical human IgG3 antibody showing the disulfide bond pattern. There are a total of 50 Cys residues and 25 disulfide bonds (–S–S–). The red parts are the variable (V) regions and the black parts are the constant (C) regions. H and L indicate the heavy and light chains, respectively; VL and CL are domains of the light chain; VH, CH1, CH2, and CH3 are domains of the heavy chain. The hinge region has a 15-residue segment that is repeated three times. (B) Expected tryptic dipeptides from human IgG3 constant region.

We verified the presence of these expected disulfide bonds prior to investigating disulfide bond variants. For example, Figure 3 shows XIC's and ETD data that support the assignment of the disulfide bonded peptide in the CH2 domain of the Fitzgerald IgG3 antibody. The XIC's of the CH2-1 (TPEVTCVVVDVSHEDPEVQFK) and CH2-2 (CK) Cys-containing peptides were plotted by searching ETD data in the *m/z* range of 1178–1181 (which encompasses the CH2-1 theoretical *m/z* of 1179 in the plus two charge state) and 248–251 (which encompasses the CH2-

2 theoretical *m/z* of 250 in the plus one charge state), respectively. This resulted in XIC's with intense peaks at the retention time of 37.5 min for both the CH2-1 peptide (Figure 3a) and the CH2-2 peptide (Figure 3b). The presence of peaks at the same retention time in both XIC's suggests that the two peptides are potential disulfide-bonded partners. To verify if they are connected, the ETD spectrum of the ion that eluted at 37.5 min was extracted (Figure 3c). The marker ion peaks for the CH2-1 and the CH2-2 peptides (at *m/z* 1179 and 250, respectively) were conspicuously present in the ETD spectrum, and *c* and *z* ions from the CH2-1 peptide, as well as *c* and *z* ions containing the intact disulfide bond (ions labeled in red) were identified, thereby confirming that the two peptides are indeed linked by a disulfide bond. Additionally, using high resolution data, the monoisotopic molecular mass of the precursor ion was determined to be within 3 ppm of the theoretical mass of the dipeptide. Hence the CH2 domain disulfide bond was assigned. The CH3, kappa and lambda constant light chain (CL), and the kappa and lambda heavy chain-light chain (HC-LC) domain disulfide bonds were assigned in a similar manner. Assignment of the lambda light chain disulfide bond is shown in Figure 4.

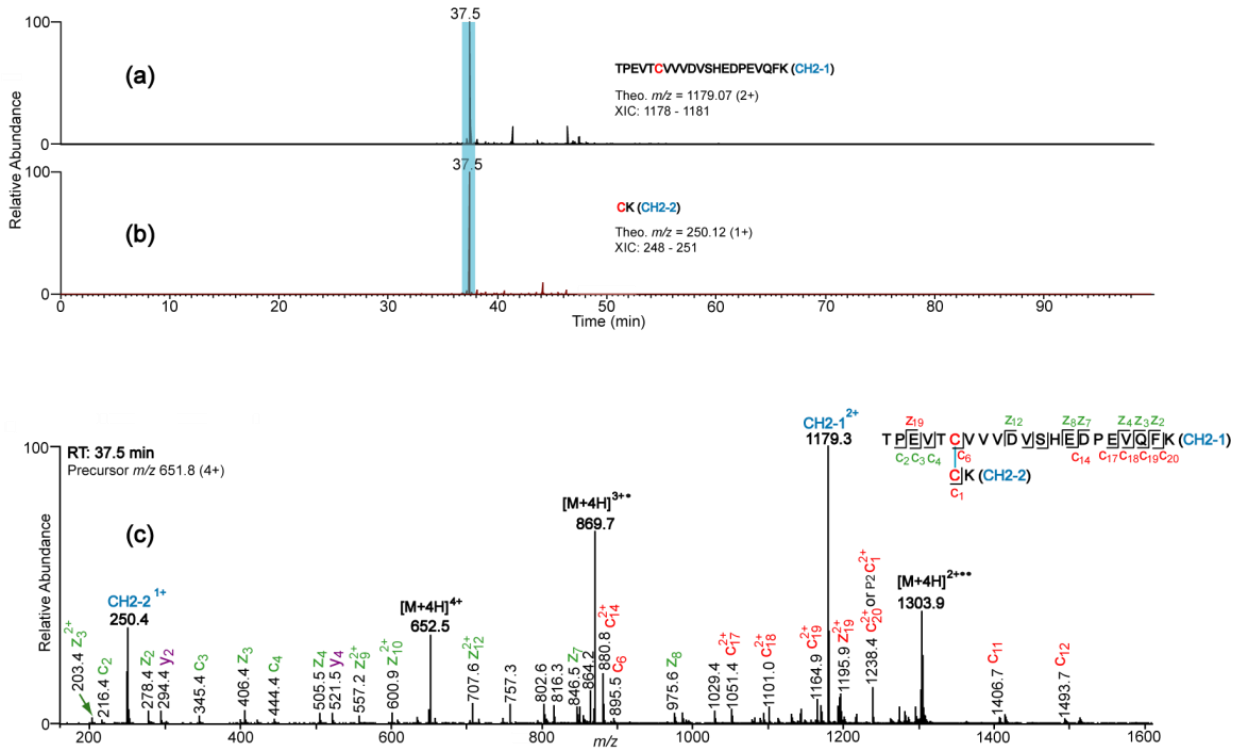


Figure 3. XIC's and ETD spectrum showing assignment of the CH2 domain disulfide bond. **(a)** and **(b)** are XIC's of peptides TPEVTCVVVDVSHEDPEVQFK and CK, respectively; clearly showing the RT of the dipeptide (highlighted). The peptide sequence, theoretical *m/z*, and the *m/z* range that was used to plot the XIC's are shown in the inset. **(c)** ETD spectrum of the CH2 domain dipeptide. Marker ion peaks resulting from the cleavage of the disulfide bond are labeled in blue; product ions (c/z ions) not containing the disulfide bond are labeled in green; product ions containing the disulfide bond are labeled in red; unexpected product ions (b/y ions) are labeled in purple.

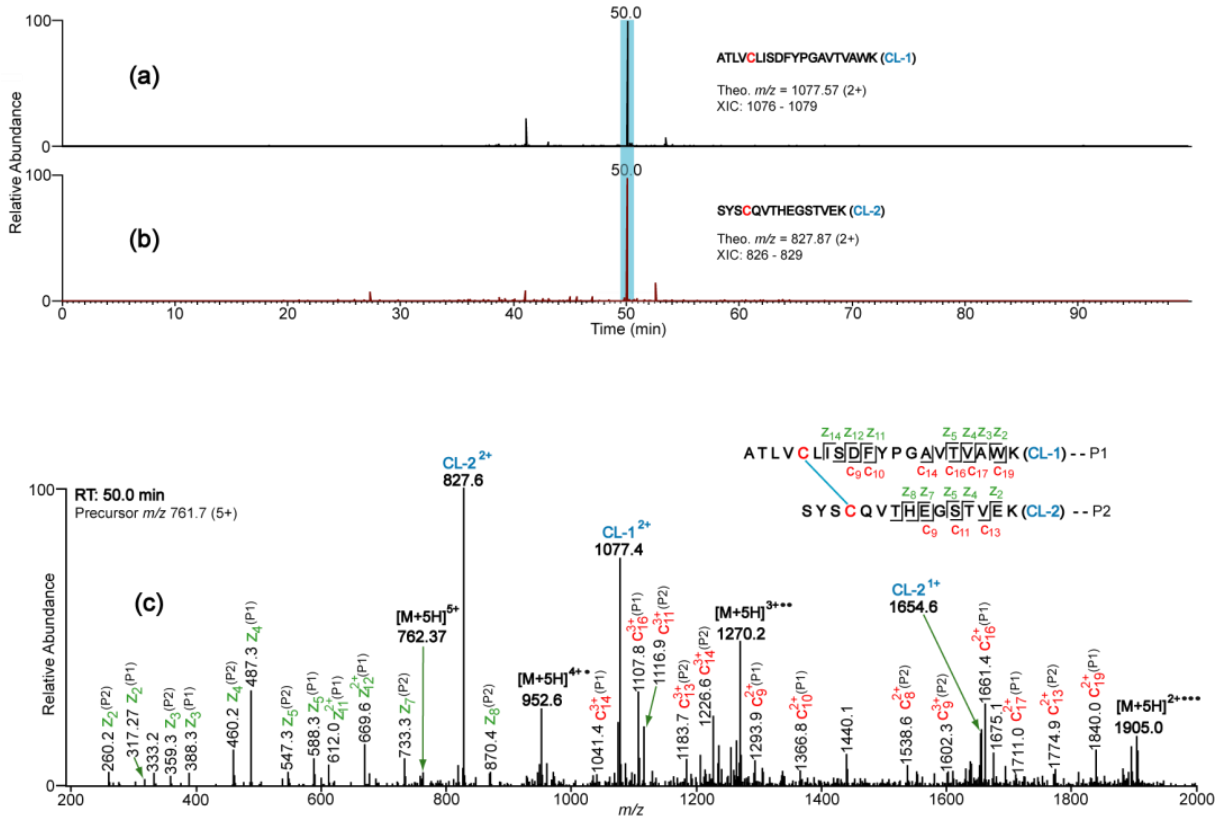


Figure 4. XIC's (**a** and **b**) and ETD spectrum (**c**) showing assignment of the classical disulfide bond in the constant region of IgG3 lambda light chain. See Figure 3 for color codes.

2.3.3 Assignment of Disulfide Bonds between Identical Cys-containing Peptides

Proteolytic digestion of IgG antibodies usually produces dipeptides with identical disulfide-bonded chains originating from the hinge regions of the antibodies. This happens because all IgG antibodies have inter-chain disulfide bonds in their hinge regions that link identical sequences of the heavy chains. For instance, IgG3 has four tryptic dipeptides (Hinge-1, Hinge-2, Hinge-3, and Hinge-4) that have identical disulfide bonded chains. Our previously published method for assigning disulfide bonds does not address this possibility; therefore, we have improved the method to account for these species. Because the disulfide bonded peptides

are identical, it may seem that the XIC data would not be useful for identifying them, since plotting the same XIC data twice results in two identical chromatograms. Nonetheless, we determined that the XIC's do, in fact, become useful if the charge states of the peptide marker ions are different for the two XIC's. For example, Figure 5 shows the assignment of the Hinge-1 disulfide bond between two identical tryptic peptides (TPLGDTTHTCPR disulfide-linked to TPLGDTTHTCPR). Although the disulfide-bonded peptides are identical, by using charge states of 1 and 2 (corresponding to *m/z* 1298.6 and 649.8, respectively) distinct XIC's were obtained (Figure 5a and b), and the peaks in the two XIC's with the same retention time correspond to the disulfide-bonded dimers. The ETD spectrum of the XIC peak at RT 22.1 min (Figure 5c) shows the marker ion peaks at *m/z* 649 and 1298, along with *c* and *z* product ion peaks with and without the disulfide bond, labeled in red and green respectively. Additionally, the observed monoisotopic mass of the precursor ion was within 2 ppm of the theoretical mass of the expected Hinge-1 dipeptide, further confirming the assignment. The Hinge-4 disulfide-linked peptide, which also contains identical Cys-containing peptides linked by a disulfide bond, was assigned in a similar manner.

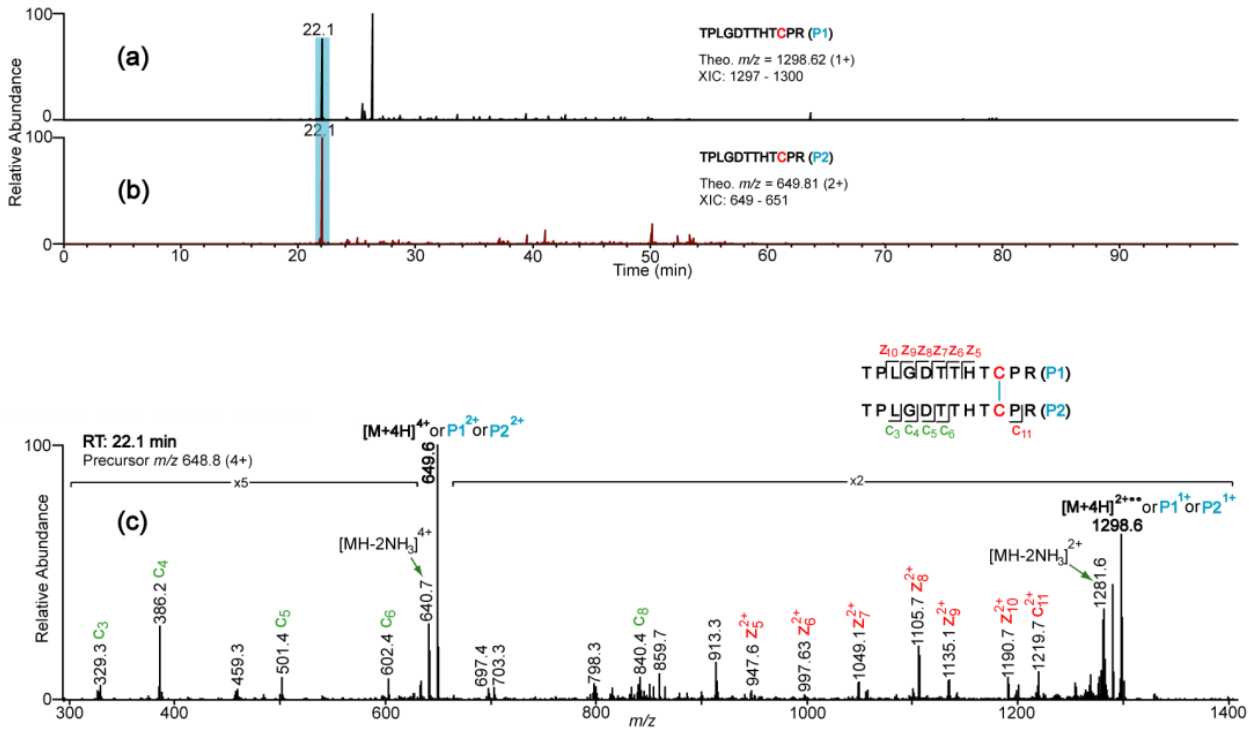


Figure 5. Representative XIC's (**a** and **b**) and ETD spectrum (**c**) that support the assignment of the Hinge-1 disulfide which has identical Cys-containing peptides. Since the peptides are identical, different charge states (+1 and +2) were used to plot their XIC's so as to avoid plotting the same XIC twice. Details about the fragment ion colors are given in Figure 3.

2.3.4 Disulfide Bonds Identified using CID Data

Three expected disulfide-bonded peptides (in the Hinge-2, Hinge-3, and CH1 regions) were not readily assigned using the XIC/ETD method. The Hinge-2 dipeptide is small (CPEPK bonded to CPEPK), and the highest charge state identified for the dipeptide was two, which is not sufficient for efficient ETD fragmentation. Additionally, the presence of two proline residues prevented ETD cleavage before and after the proline residues, limiting the formation of *c* and *z* ions. The Hinge-3 dipeptide, which was also not rapidly detected using the XIC/ ETD method, contains two SCDTPPPCPR tryptic peptides connected by two disulfide bonds. In this case,

ETD data was acquired on the dipeptide, but the ETD spectrum showed only charged reduced species and no peptide marker ions. Therefore, the XIC's of the disulfide-bonded partners were not useful for assigning this dipeptide. The disulfide-bonded peptide in the CH1 domain also was not identified by the XIC/ ETD method because XIC's of these tryptic peptides did not show marker ions for the two peptide partners in the same ETD spectrum. The absence of these ions could be due to the large size of one of the tryptic peptides. The CH1-2 tryptic peptide is 63 amino acids long (without any missed cleavage) and the Cys residue is 11 residues from the C-terminus, leaving 52 amino acids after the disulfide bond. It is possible that because the portion of the peptide is very long, it could fold around the disulfide bond, thereby preventing efficient transfer of the ETD reagent ion to the disulfide bond and consequent cleavage of the bond by ETD. A shorter CH1 domain dipeptide may be obtainable using a different enzyme; in that case, the XIC/ETD method may be used to assign the disulfide bond.

These three disulfide-linked peptides were assigned using a complementary strategy: searching for the in-tact disulfide-bonded peptides in the high resolution MS data first, followed by confirmation of the species using CID data, a procedure reported by Go *et al*.[22] A prediction table containing the theoretical masses and *m/z's* of the dipeptides at different charge states was generated, and the XIC of each intact dipeptide was constructed from the total ion chromatogram. High resolution MS data and the corresponding CID spectra were used to assign the dipeptides. For instance, Figure 6 shows the assignment of the CH1 domain disulfide bond using this approach. The XIC of the ion at *m/z* 1130 was plotted (Figure 6A) and the corresponding high resolution mass spectrum at 53.3 min is shown in the insert. It contains five peaks at *m/z* 879, 989, 1130, 1319, and 1582, which correspond to the theoretical *m/z's* values of

the CH1 domain dipeptide at charge states of 9+, 8+, 7+, 6+, and 5+, respectively. Figure 6B shows the CID spectrum of the ion, *m/z* 1130, eluting at 53.3 min. Abundant b and y ions resulting from the fragmentation of both CH1 bonded peptides are present, and they are used to unequivocally confirm the CH1 disulfide-bonded peptide. This data analysis approach was also used for the assignment of the Hinge-2 and Hinge-3 disulfide-bonded peptides

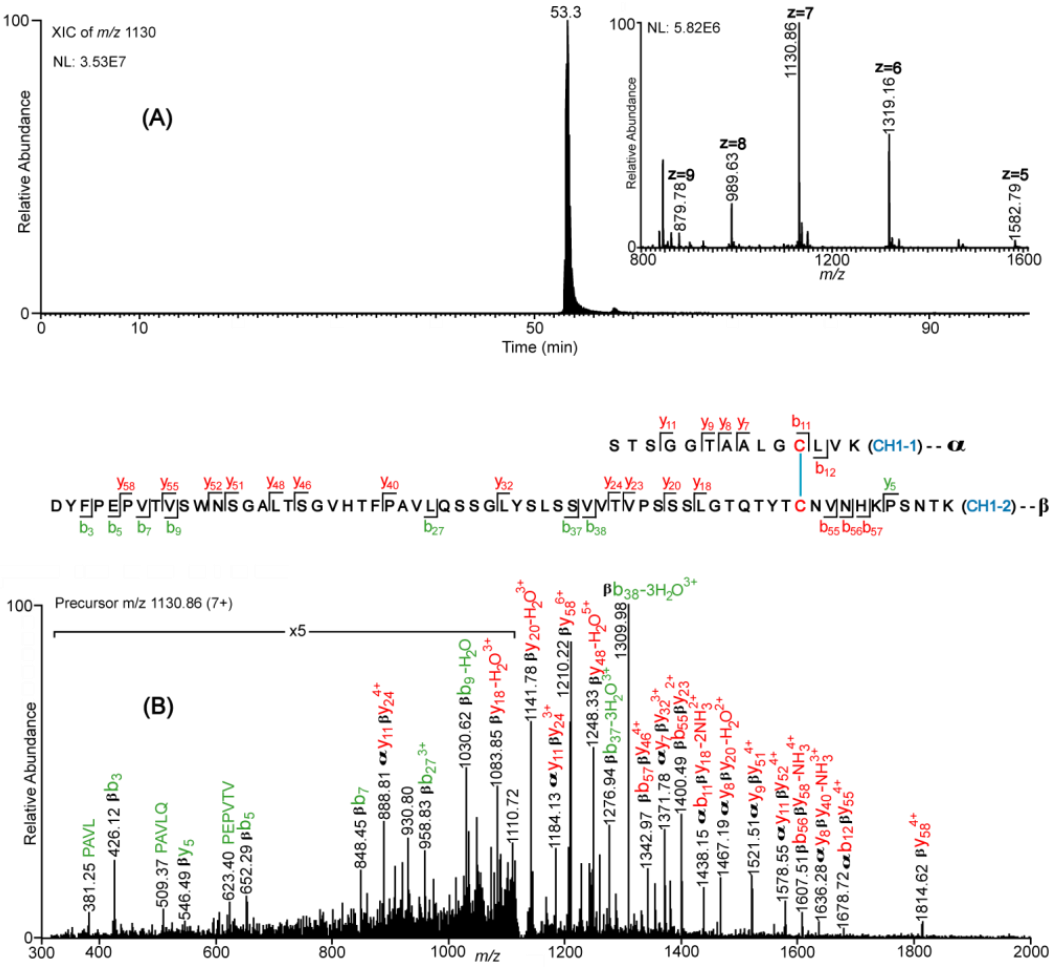


Figure 6. XIC and CID spectrum supporting the assignment of the CH1 domain disulfide bond. **(A)** XIC of the CH1 domain tryptic dipeptide at *m/z* 1130 (7+). The insert shows the MS^1 spectrum extracted from the XIC peak at 53.3 min. **(B)** CID spectrum of the 1130.86 ion (7+), supporting the assignment of the CH1 domain disulfide bond. Fragment ions (b/y) that contain the disulfide bond are labeled in red; fragment ions that do not contain the disulfide bond are labeled in green.

Overall, by combining two different MS-based approaches for assigning disulfide-linked peptides,[20,22] we identified all the expected disulfide bonds in the constant region of endogenous IgG3 from two different sources. A summary of the disulfide bond assignments using the XIC/ETD method is shown in Table 1, and those identified using high resolution and CID data are shown in Table 2.

Table 1. Summary of Disulfide Bond Assignments in Fitzgerald and Sigma IgG3 mAbs Using XIC/ETD Data.

			Fitzgerald IgG3		Sigma IgG3	
Cys-Containing Tryptic peptides	Position	Peptide XIC m/z (Theo)	RT	Peptide XIC m/z (Expt)	RT	Peptide XIC m/z (Expt)
GPSVFPLAPCSR	HC-LC	1230.6	34.9	1229.7	35.1	1229.7
TVAPTECS		807.4	34.9	807.4	35.1	807.4
ATLVCLISDFYPGAVTVAWK	CL	1077.6^{2+}	50.0	1077.4^{2+}	50.3	1077.4^{2+}
SYSCQVTHEGSTVEK		827.9^{2+}	50.0	827.6^{2+}	50.3	827.7^{2+}
TPLGDTTHTCPR	Hinge-1	1298.6	22.1	1298.6	22.6	1298.8
TPLGDTTHTCPR		649.8^{2+}	22.1	649.6^{2+}	22.6	649.6^{2+}
CPAPELLGGPSVFLFPPKPK	Hinge-4	1047.6^{2+}	53.4	1047.8^{2+}	53.7	1047.8^{2+}
CPAPELLGGPSVFLFPPKPK		698.7^{3+}	53.4	698.9	53.7	698.8^{3+}
TPEVTCVVVDVSHEDPEVQFK	CH2	1179.1^{2+}	37.5	1179.3^{2+}	37.8	1179.6^{2+}
CK		250.1	37.5	250.4	37.8	250.1
WQQGNIFSCSVMHEALHNR	CH3	1129.0^{2+}	41.0	1128.9^{2+}	41.2	1129.2^{2+}
NQVSLTCLVK		1104.6	41.0	1104.7	41.2	1104.6

Unless otherwise stated, all ions are in the plus one charge state.

Table 2. Summary of Mass Assignments of Disulfide Bonds in Fitzgerald and Sigma IgG3 mAbs Using High Resolution Data.

Tryptic dipeptides	Position	Theor. m/z	CS	Fitzgerald IgG3		Sigma IgG3	
				Exptl. m/z	Mass Error (ppm)	Exptl. m/z	Mass Error (ppm)
DYFPEPVTVSWNSGALTS GVHTFPAVLQSSGLYSLSS VVTVPSSSLGTQTYTCNVN HKPSNTK	CH1	1581.9839	5+	1581.9953	5	1581.9974	9
		1318.4878	6+	1318.4954	6	1318.4954	5
		1130.2763	7+	1130.2854	8	1130.2886	11
		989.1177	8+	989.1251	7	989.1251	8
STSGGTAALGCLVK		879.3276	9+	879.3309	4	879.3309	8
CPEPK	Hinge-2	572.2623	2+	572.2564	5	572.2602	4
CPEPK							
SCDTPPPCPR	Hinge-3	1070.4394	2+	1070.4419	2	1070.4434	4
		713.9620	3+	713.9645	4	713.9641	< 1
SCDTPPPCPR		535.7234	4+	535.7264	6	535.7250	3

2.3.5 Assignment of Disulfide Bond Variants (Alternative Disulfide Bonds)

After verification that all of the classical IgG3 disulfide bonds were present, we investigated the presence of alternative disulfide bonds. The same approaches used to identify the classical disulfide bonds were used to search for alternative disulfide bonds. For the dipeptides that were identified using the XIC/ ETD method, the XIC's of all the Cys-containing peptides were aligned and compared (Figure 7). In addition to the peaks that revealed the expected disulfide bonds (highlighted in blue), three sets of low abundant peaks in the XIC's of peptides that are not expected to be disulfide bonded were identified to have the same retention time (peaks at 19.2, 38.1 and 41.0 min), suggesting the possible presence of alternative disulfide bonds. To determine if the ions generating these peaks are aberrant disulfide-bonded peptides,

the corresponding ETD spectra were extracted and studied. None was found to correspond to disulfide-bonded peptides. For example, the XIC's of CPAPELLGGPSVFLFPPKPK (Hinge-4, P2, *m/z* 698.72 at charge state of three) and WQQGNIFSCSVMHEALHNR (CH3-1, *m/z* 1129.02 at charge state of two), which are not expected to be disulfide bonded, both have peaks at 19.2 min, suggesting that they are potentially linked by an alternative disulfide bond. The MS data that generated the ETD spectrum of the ion that eluted at 19.2 min was interrogated. Immediately, this peak was confirmed not to be an aberrantly disulfide-bonded peptide because the precursor ion mass that generated this peak was from a 1267.6 Da ion, which does not match the theoretical mass of 4347.2 for disulfide-bonded CPAPELLGGPSVFLFPPKPK and WQQGNIFSCSVMHEALHNR. Hence, there was no alternative bond between these two peptides. After further interrogation, this particular peptide was assigned as the Hinge-2 dipeptide with a non-specific N-ethylmaleimide alkylation (theoretical mass of 1267.6). The MS data for the peaks at 38.1 and 41.0 min were interrogated in a similar manner, and they were also found not be related to any alternatively linked peptides. Overall, no alternative disulfide bonded peptides were found between any Cys-containing peptides that are not expected to be disulfide-bonded.

For the CH1, Hinge-2, and Hinge-3 domain disulfide bonds (see Figure 2B for the tryptic disulfide-bonded peptides), which were identified using high resolution (MS^1) and CID data, a prediction table containing the masses and *m/z* values of plausible alternative disulfide bonds between these Cys-containing peptides and other Cys-containing peptides in close proximity was constructed, as described previously.[22] This table includes six plausible four-peptide disulfide-linked chains involving two SCDTPPPCPR peptides and three other Cys-containing peptides in

close proximity. The calculated *m/z's* were searched using high resolution MS and CID data, and no alternative disulfide-linked peptides were identified in either IgG3 sample using this approach.

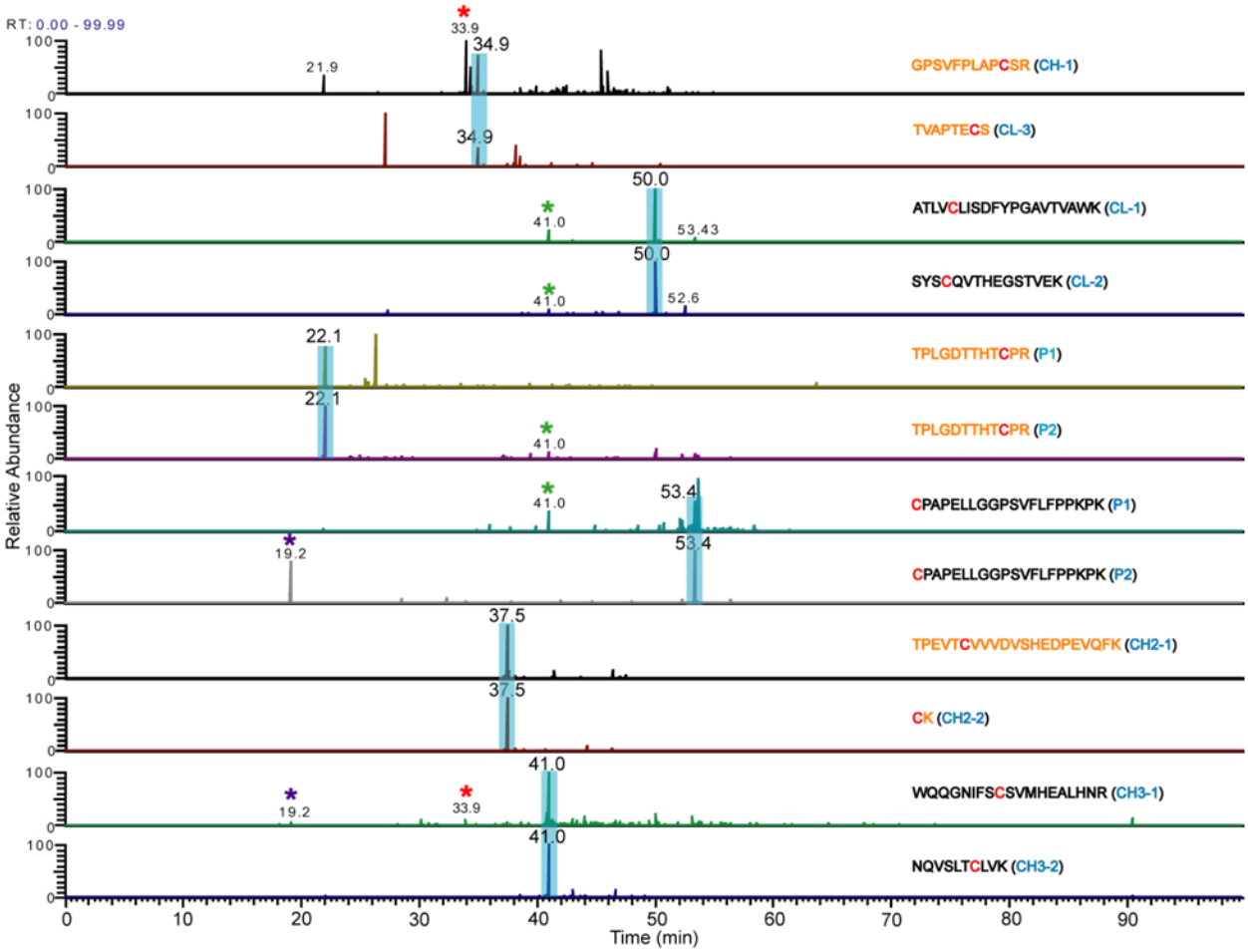


Figure 7. XIC's of Cys-containing peptides of the disulfides identified using the XIC/ETD method. Adjacent Cys-containing peptides of the same color (orange or black) are expected to be disulfide bonded and the peaks in their XIC's that lead to the identification of the dipeptide are highlighted in blue. Peaks having the same retention time in XIC's of peptides that are not expected to be disulfide bonded are indicated by asterisks of the same color. The peaks at 19.2 and 33.9 were interferences since they did not correspond to any dipeptide, while the peaks at 41.0 were the CH3 disulfide that showed up as interference in some other XIC's.

In summary, no disulfide bond variants were detected, even though they were searched for using two different search strategies. It is theoretically possible that disulfide bond variants exist for IgG3 and remained undetected, but if that is the case, we expect these variants to be in very low abundance, perhaps less than 2% of the protein. This estimate is based on substantial prior work we have completed using these methods and mapping disulfide bond variants in HIV-1 envelope proteins.[20,22–26] In one case, the MS methods described here were able to detect an aberrant disulfide-bonded isoform that was present in substantially less than 5% of the total protein population.[26]

2.4 Conclusion

We carried out an extensive disulfide bond analysis of two endogenous human IgG3 samples to experimentally verify the presence of the classical IgG3 disulfide bonds and to investigate the possibility of disulfide-mediated isoforms resulting from alternative disulfide bonds. All expected disulfide-bonded peptides in the constant regions of the IgG3 samples from two different sources were unambiguously assigned. Both the kappa and lambda forms were fully characterized in each protein. Although disulfide-mediated isoforms have been identified for IgG2 and IgG4 antibodies, and to a lesser extent in IgG1, the data presented herein show that there are no alternative disulfide bonded peptides within the constant region of native IgG3 antibodies, indicating that endogenous IgG3 antibodies do not have disulfide-mediated isoforms. These data provide the first benchmark for a complete native IgG3 disulfide bonding profile, and the analytical approach described herein can be readily applied to recombinant IgG3 antibodies or any other IgG.

2.5 Acknowledgements

This manuscript is dedicated to Professor Craig E. Lunte, who was a wonderful leader and friend of the bioanalytical community. The work was supported by NIH grant number R01AI094797 and R01GM103547 to Heather Desaire and R01GM090080 to Thomas Tolbert.

2.6 References

1. Bruhns, P.; Iannascoli, B.; England, P.; Mancardi, D. A.; Fernandez, N.; Jorieux, S.; Daeron, M. *Blood,* **2009,** 113, 3716-3725.

2. Hogarth, P. M.; Pietersz, G. A. *Nat. Rev. Drug Discov.*, **2012,** 11, 311-331.

3. Morell, A.; Terry, W. D.; Waldmann, T. A. *J. Clin. Invest.,* **1970,** 49, 673-680.

4. Jefferis, R. *Expert Opin. Biol. Ther.,* **2007,** 7, 1401-1413.

5. Stapleton, N. M.; Andersen, J. T.; Stemerding, A. M.; Bjarnarson, S. P.; Verheul, R. C.; Gerritsen, J.; Zhao, Y.; Kleijer, M.; Sandlie, I.; de Haas, M.; Jonsdottir, I.; van der Schoot, C. E.; Vidarsson, G. *Nat. Commun.,* **2011,** *2*:599.

6. Plomp, R.; Dekkers, G.; Rombouts, Y.; Visser, R.; Koeleman, C. A. M.; Kammeijer, G. S. M.; Jansen, B. C.; Rispens, T.; Hensbergen, P. J.; Vidarsson, G.; Wuhrer, M. *Mol. Cell. Proteomics,* **2015,** 14, 1373-1384.

7. Abkevich, V. I.; Shakhnovich, E. I. *J. Mol. Biol.,* **2000,** 300, 975-985.

8. Zhang, L.; Chou, C. P.; Moo-Young, M. *Biotechnol. Adv.,* **2011,** 29, 923-929.

9. Betz, S. F. *Protein Sci.,* **1993,** 2, 1551-1558.

10. Milstein, C.; Frangion. B. *Biochem. J.,* **1971,** 121, 217-225.

11. Frangione, B.; Milstein, C. *J. Mol. Biol.,* **1968,** 33, 893-906.

12. Frangione, B.; Milstein, C. *Nature,* **1967,** 216, 939-941.

13. Milstein, J. R. P. a. C. *Nature*, **1967,** 216, 941-942.

14. Pink, J. R.; Milstein, C. *Nature,* **1967,** 214 (5083), 92-94.

15. Schuurman, J.; Perdok, G. J.; Gorter, A. D.; Aalberse, R. C. *Mol. Immunol.,* **2001,** 38, 1-8.

16. Bloom, J. W.; Madanat, M. S.; Marriott, D.; Wong, T.; Chan, S. Y. *Protein Sci.,* **1997,** 6, 407-415.

17. Wypych, J.; Li, M.; Guo, A.; Zhang, Z.; Martinez, T.; Allen, M. J.; Fodor, S.; Kelner, D. N.; Flynn, G. C.; Liu, Y. D.; Bondarenko, P. V.; Ricci, M. S.; Dillon, T. M.; Balland, A. *J. Biol. Chem.,* **2008,** 283, 16194-16205.

18. Martinez, T.; Guo, A.; Allen, M. J.; Han, M.; Pace, D.; Jones, J.; Gillespie, R.; Ketchem, R. R.; Zhang, Y.; Balland, A. *Biochemistry,* **2008,** 47, 7496-7508.

19. Zhang, A.; Fang, J.; Chou, R. Y. T.; Bondarenko, P. V.; Zhang, Z. *Biochemistry,* **2015,** 54, 1956-1962.

20. Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.,* **2013,** 85, 1192-1199.

21. Wu, S.-L.; Jiang, H.; Lu, Q.; Dai, S.; Hancock, W.; Karger, B. L. *Anal. Chem.,* **2009,** 81, 112-122.

22. Go, E. P.; Hua, D.; Desaire, H. *J. Proteome Res.,* **2014,** 13, 4012-4027.

23. Eden P. Go , Y. Z., Sushma Menon.; Desaire, H. *J. Proteome Res.,* **2011,** 10, 578-591.

24. Kassa, A.; Dey, A. K.; Sarkar, P.; Labranche, C.; Go, E. P.; Clark, D. F.; Sun, Y.; Nandi, A.; Hartog, K.; Desaire, H.; Montefiori, D.; Carfi, A.; Srivastava, I. K.; Barnett, S. W. *PloS one,* **2013,** 8, e76139.

25. Ringe, R. P.; Yasmeen, A.; Ozorowski, G.; Go, E. P.; Pritchard, L. K.; Guttman, M.;

Ketas, T. A.; Cottrell, C. A.; Wilson, I. A.; Sanders, R. W.; Cupo, A.; Crispin, M.; Lee, K. K.; Desaire, H.; Ward, A. B.; Klasse, P. J.; Moore, J. P. *J Virol.,* **2015,** 89, 12189-12210.

26. Go, E. P.; Cupo, A.; Ringe, R.; Pugach, P.; Moore, J. P.; Desaire, H. *J Virol.*, **2015,** 90, 2884-2894.

Chapter 3: Two New Tools for Glycopeptide Analysis Researchers: a Glycopeptide Decoy Generator and a Large Dataset of Assigned CID Spectra of Glycopeptides

*The work described in this chapter encompasses an original publication in the Journal of Proteome Research: Jude C. Lakbub, Xiaomeng Su, Zhikai Zhu, Milani W. Patabandige, David Hua, Eden P. Go, Heather Desaire. J. Proteome Res., **2017**, 16 (8), 3002-3008. It is reproduced here with permission from The American Chemical Society.*

Abstract

The glycopeptide analysis field is tightly constrained by a lack of effective tools that translate mass spectrometry data into meaningful chemical information, and perhaps the most challenging aspect of building effective glycopeptide analysis software is designing an accurate scoring algorithm for MS/MS data. Herein, we provide the glycoproteomics community with two tools to address this challenge. The first tool, a curated set of 100 expert-assigned CID spectra of glycopeptides, contains a diverse set of spectra from a variety of glycan types; the second tool, Glycopeptide Decoy Generator, is a new software application that generates glycopeptide decoys *de novo*. We developed these tools so that emerging methods of assigning glycopeptides' CID spectra could be rigorously tested. Software developers or those interested in developing skills in expert (manual) analysis can use these tools to facilitate their work. We demonstrate the tools' utility in assessing the quality of one particular glycopeptide software package, GlycoPep Grader, which assigns glycopeptides to CID spectra. We first acquired the set of 100 expert assigned CID spectra; then we used the Decoy Generator (described herein) to generate 20 decoys per target glycopeptide. The assigned spectra and decoys were used to test

the accuracy of GlycoPep Grader's scoring algorithm; new strengths and weaknesses were identified in the algorithm using this approach. Both newly-developed tools are freely available to interested parties. The software can be downloaded at http://glycopro.chem.ku.edu/GPJ.jar.

3.1 Introduction

Glycosylation is a common but complex post-translational modification that occurs on proteins during their biosynthesis, and it is known to regulate several biological processes such as cell signaling,[1,2] protein folding,[3,4] transportation,[3,5] and degradation.[5] Changes in the glycosylation profiles of endogenous glycoproteins can serve as biomarkers for diseases diagnosis and progression.[6,7] In addition, glycosylation can impact the biological activity,[8,9] immunogenicity,[9,10] and stability[11] of glycoprotein-based drugs. Hence, extensive characterization of glycosylation on glycoproteins is vital in understanding important biological events and diseases, as well as the pharmacodynamics and pharmacokinetics of glycoprotein-based drugs.

Mass spectrometry (MS) has become an invaluable analytical tool for glycosylation characterization due to its high sensitivity, high resolution, and complementary fragmentation techniques.[12] Two main methods for mass spectrometric glycosylation analysis of glycoproteins are the glycan-based approach[13,14] and the glycopeptide-based approach.[15,16] While the former approach gives information about the total glycan pool on a glycoprotein, the latter approach provides glycosylation site-specific information. Because glycopeptide analysis is the method of choice for glycosylation profiling of proteins containing more than one glycosylation site, we focus on it herein. Although advances in mass spectrometry instrumentation, sample preparation,[17,18] and data acquisition methods[19,20] have contributed to advances in glycopeptide

analyses, interpretation of the resulting mass spectrometry data from tandem MS experiments remains an additional ongoing field of development.

An area of increasing interest in glycopeptide analysis is, therefore, the development of bioinformatics tools for rapid and automated assignment of glycopeptides to MS/MS data. Glycopeptides are typically analyzed by tandem mass spectrometry using fragmentation methods such as collision induced dissociation (CID), electron transfer dissociation (ETD), and higher energy collision dissociation (HCD). Manual analysis of glycopeptide data generated by these fragmentation methods provides the most confident glycopeptide assignments, but it is extremely time-consuming and requires extensive experience in data analysis. Hence, several bioinformatics tools have been developed for the interpretation of glycopeptide MS/MS data. Examples include GlycoPep Grader,[21] GlycoPeptideSearch,[22] and MAGIC[23] that assign glycopeptides to CID spectra; GPQuest[24] and pGlyco[25] for assignment of glycopeptides to HCD spectra; and GlycoPep Detector[26] and GlycoPep Evaluator[27] for assignment of glycopeptides to ETD spectra. Other tools such as GlycoFragwork[28] and GlycoMaster DB[29] assign glycopeptide spectra based on a combination of two or more fragmentation techniques. A number of reviews that describe these tools, and others, in detail have been reported.[30-32] In general, bioinformatics tools match glycopeptides to MS/MS data by scoring potential glycopeptide candidates against a particular MS/MS spectrum, and the candidate with the highest score is assigned to the spectrum.[21,26,27] However, automated glycopeptide assignments can be problematic, as the best match for a spectrum can sometimes be an incorrect match. Therefore, it is vital for researchers to assess the accuracy of algorithms that assign glycopeptides to MS/MS data in order to ensure

confidence of the results; this rigorous testing also affords developers valuable information that can be used to improve the algorithms.

Herein, we release two new bioinformatics tools to the community; they support glycopeptide analysis software developers and those assigning CID spectra of glycopeptides, either manually, or by an automated tool. The first product, Glycopeptide Decoy Generator (GDG), rapidly generates abundant decoy glycopeptides *de novo*, and enables determination of the accuracy of tools that assign glycopeptides to CID data. Large numbers of decoys can be easily generated for glycopeptides using our tool. GDG generates abundant decoys for any target glycopeptide, and all the decoys have biologically relevant glycan components. The second product we provide herein is a dataset of 100 expert-assigned CID spectra of a diverse set of glycosylated peptides. The dataset contains all major N-glycosylation types, including sialylated and fucosylated glycoforms. We demonstrate the tools' utility in assessing Glycopep Grader's scoring algorithm; this tool assigns glycopeptides to CID spectra. Both newly-developed tools described herein are freely available to interested parties.

3.2 Experimental

3.2.1 Materials and Reagents

Avidin, IgG1, bovine ribonuclease B (RNAse B), bovine fetuin, human apo-transferrin, Tris(hydroxylmethyl)aminomethane (Trizma) base, urea, dithiotreitol (DTT), iodoacetamide (IAM), and formic acid were purchased from Sigma-Aldrich (St. Louis, MO). IgG2 and IgG3 were from Fitzgerald (Acton, MA), and sequencing grade trypsin was from Promega (Madison, WI). HIV-1 envelope glycoprotein samples, C.97ZA012 gp140 and A244-V1V2, were from the

Duke Human Vaccine Institute (Durham, NC). Ultrapure water was obtained via a Direct-Q water purification system (MilliporeSigma, Darmstadt, Germany).

3.2.2 Sample Preparation

For the HIV-1 Env proteins, the samples were prepared as reported in Reference 36. Briefly, about 100 µg of each of the proteins were dissolved in 100 mM Tris buffer (pH 8.0), and urea was added to a final concentration of 6 M to denature the proteins. Subsequently, DTT was added to reduce the disulfide bonds, and the reaction was allowed to proceed for 1 h at room temperature. After disulfide bond reduction, IAM was added to a final concentration of 10 mM and incubated for 1 h in the dark to cap free cysteine residues; followed by addition of excess DTT to react with excess IAM. The excess salt (urea and DTT) were either diluted to less than 1M (for gp140) or removed by centrifugal filtration of the samples using 10 kDa molecular weight cut-off filters (for A244-V1V2), and the samples were reconstituted in Tris buffer (pH 8.0) to a final concentration of 2 µg/µL. Finally, trypsin was added at an enzyme-to-protein ratio of 1:30 (w/w) and incubated for 18 h at 37 °C, followed by a second trypsin addition at an enzyme-to-protein ratio of 1:30 (w/w) under the same conditions.

For all other glycoproteins, the sample preparation was the same as described above, but with a slight modification during the digestion step. Trypsin digestion was done at 1:30 (w/w) enzyme-to-protein ratio for 18 h at 37 °C, followed by a second trypsin addition at an enzyme-to-protein ratio of 1:100 (w/w) for additional 3 h. After digestion, all samples were quenched by addition of 1 µL formic acid for every 100 µL of sample. The samples were analyzed immediately after digestion and/or aliquoted and stored at -20 C until analysis.

3.2.3 LC Separation and MS Data Acquisition

LC-MS analysis was conducted on a Waters Acquity Ultra Performance Liquid Chromatography instrument (Waters Acquity, Milford, MA) coupled to either a LTQ Velos Linear Ion Trap or a LTQ Orbitrap Velos Pro hybrid Mass Spectrometer (Thermo Scientific, San Jose, CA). For avidin and the HIV-1 Env protein, C.97ZA012 gp140, data was acquired on the LTQ Velos Linear Ion Trap, and the column dimensions, gradient, and CID data acquisition settings are the same as those reported in Reference 27. For the remaining proteins, data was acquired on the LTQ Orbitrap Velos Pro. The Mobile Phase A was 99.9% LC/MS-grade water containing 0.1% formic acid and Mobile Phase B was 99.9% acetonitrile with 0.1% formic acid. A C18 Aquasil Gold column (100 x 1 mm i.d, 175 Å, Thermos Scientific, San Jose, CA) was used for reversed phase separation. Sample solutions of 5 µL were injected onto the column and separated at a flow rate of 50 µL/min as follows: The mobile phase B was initially maintained at 2% for 5 min followed by an increase to 35% in 60 min, and then ramped to 60% in 15 min. Mobile phase B was held at 95% for 10 min prior to re-equilibration of the column at 2% B for 10 min.

For both the LTQ Velos and the Orbitrap Velos Pro mass spectrometers, data-dependent acquisition was performed in the positive ion mode, and the acquisition parameters were optimized for each protein. The ESI source spray voltage was maintained at 3.0 kV and the capillary temperature was 200 °C for HIV-1 C.97ZA012 gp140, 260 °C for RNAse B, 275 °C for IgG2, and 250 °C for all other proteins. In all experiments, a survey MS scan was obtained from *m/z* 400 or 500 to 2000 prior to CID fragmentation in the linear ion trap. For MS scans obtained in the Orbitrap mass analyzer, the resolution was set at 30,000 (at *m/z* 400). CID spectra were

obtained by selecting the top 5 ions (top 8 for RNAse B) for CID fragmentation in the linear ion trap. The CID normalization collision energy was 35% (30% for HIV-1 C.97ZA012 gp140) with an activation time of 10 ms and a 3 Da isolation window.

3.2.4 Glycopeptide Spectral Library

A library of glycopeptides' CID spectra was generated using the following proteins: IgG1, IgG2, IgG3, bovine fetuin, RNAse B, avidin, transferrin, and two HIV-1 Env proteins (C.97ZA012 gp140 and A244-V1V2). These glycoproteins, with the exception of A244-V1V2, have been well-characterized and reported in literature.[33-37] For the glycopeptides that have been reported in literature, the dataset of manually characterized CID spectra was generated as follows: For each glycoprotein, a list of previously assigned glycopeptide compositions was compiled, and the theoretical monoisotopic *m/z* values at different charge states were computed and searched for in the MS data file of the glycopeptide digest of the protein. When a match was found, we determined if the peak was selected for CID fragmentation. If a corresponding CID spectrum was found within 1 minute of the retention time of the peak in the full-MS scan, and if characteristic glycan oxonium ions (e.g ions at *m/z* 366, 528, and 690) were identified, the spectrum was manually assigned based on knowledge of glycopeptide fragmentation under CID conditions. For A244-V1V2, which has not been reported in the literature, the glycopeptides were assigned using a previously described workflow for characterizing glycosylation on complex glycoproteins.[36,38] Briefly, compositional analysis of glycopeptides was carried out by first doing an *in silico* digestion of the protein to find peptides with the N-X-S/T glycosylation site motif; then CID spectra were identified that contained an abundant ion consistent with the Y_1 ions[39,40] (ions which are typically used to identify the peptide portions of glycopeptide

compositions) that would be generated from these glycopeptides. Once candidate CID spectra were identified in this way, plausible glycopeptide compositions were obtained using high-resolution MS data and GlycoPep DB.[41] Potential glycopeptide candidates with experimental monoisotopic *m/z* values within 10 ppm from the theoretical *m/z* of the glycopeptide reported by GlycoPep DB were confirmed manually by annotating the glycosidic cleavages observed in the CID data. Overall, for all the proteins, each confirmed glycopeptide assignment in the dataset met the following criteria: (1) the experimental monoisotopic mass of the precursor ion closely matched the theoretical monoisotopic mass of the assigned glycopeptide (within 10 ppm for Orbitrap data and 30 ppm for LTQ Velos data); (2) the CID spectrum contained an intense Y_1 ion of the glycopeptide; (3) the CID spectrum contained all or some of the following characteristic oxonium ion peaks: *m/z* 366, 528, 690, and 657 (for sialylated glycopeptides); and (4) glycosidic cleavages consistent with neutral losses of the monosaccharides present in the glycopeptide were observed. The Supplemental Data includes 13 example annotated spectra. Furthermore, the peaklists for all one hundred CID spectra are supplied in the Supplemental Data, along with their assigned glycopeptide compositions. These spectral data can be used by other software developers who wish to test CID algorithms against expert-verified, pre-assigned data.

3.3 Results and Discussion

3.3.1 Tool 1: The Glycopeptide Decoy Generator

Glycopeptide Decoy Generator (GDG) is a free tool designed to generate abundant glycopeptide decoys for accurate assessment of glycopeptide scoring algorithms that match CID

spectra to glycopeptide compositions. Figure 1A shows the graphical user interface of the decoy generator. GDG contains two main menus, the “Input Data” menu and the “Result” menu.

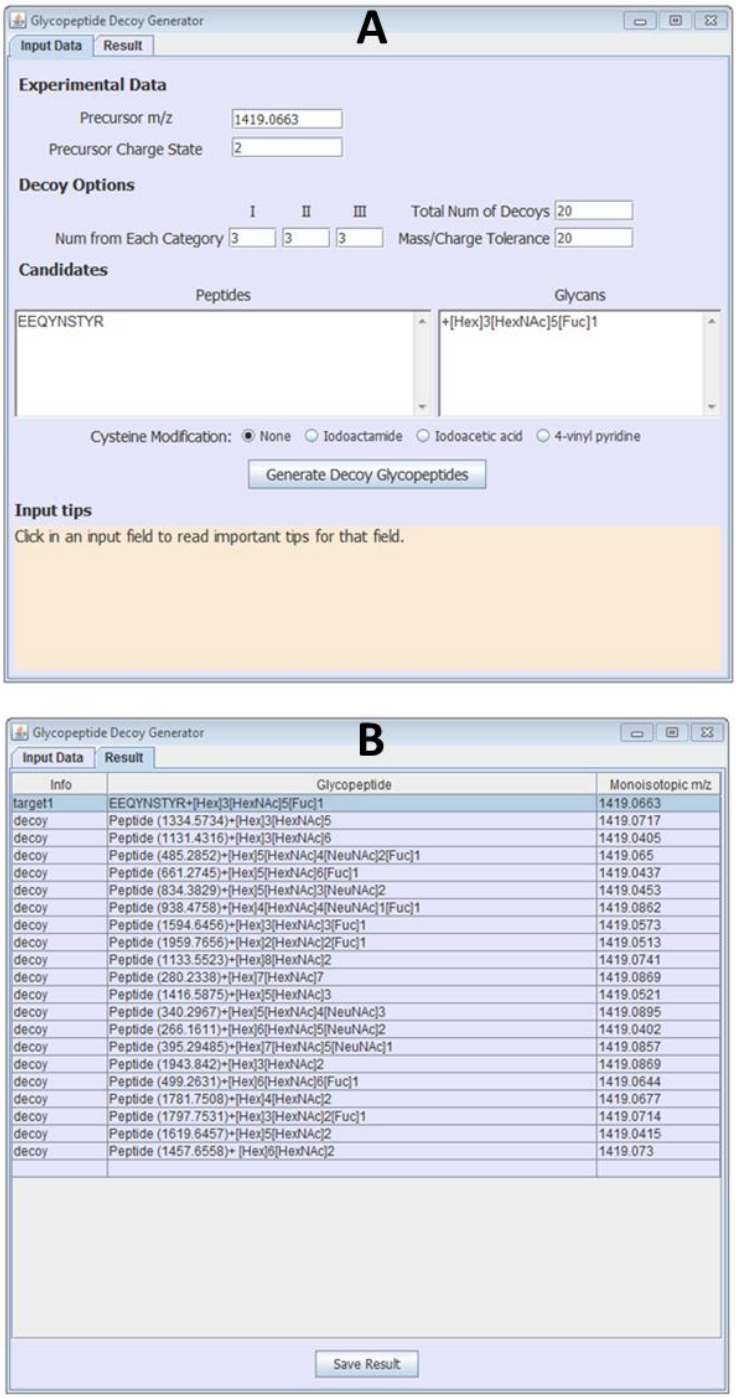


Figure 1. Graphical user interface of Glycopeptide Decoy Generator showing **(A)** the "Input Data" menu and parameters to generate 20 decoys for a target glycopeptide; and **(B)** the "Result" menu showing 20 decoys generated for the target glycopeptide in **(A)**.

To generate decoys, the user enters the monoisotopic *m/z* value and charge state of the target glycopeptide composition (target), the number of decoys to generate, as well as the desired mass tolerance (in ppm) of the decoys from the target. In addition, the peptide and glycan portions of the target glycopeptide are entered in two adjacent windows with each peptide portion aligned with its glycan portion. The decoys generated by GDG can be divided into three categories based on their glycan compositions, and the user may enter the number of decoys required for each category ("Num from Each Category"). The three decoy categories are: (I) decoys containing [HexNAc]2[Hex]1-n[Fuc]0-2, where n is any integer greater than 1; (II) decoys containing [NeuNAc], and (III) decoy glycopeptides not belonging to categories (I) or (II). If the sum of the numbers entered for each decoy category is lower than the total number of decoys entered by the user, the software randomly adds decoys from all three categories to make up the total number of decoys required. Finally, if the peptide portion contains a cysteine (Cys) residue, the user must specify whether or not the cysteine residue is modified. The current version of GDG has options for cysteine modification using iodoacetamide, iodoacetic acid, and vinyl pyridine, which are the commonly used Cys alkylating agents. For any other modification, including cysteine modification with other alkylating agents like N-ethylmaleimide, the mass of the modification can be entered in brackets after the amino acid residue that is modified. For example, if the peptide DETMFNASQR has Met oxidation, it would be entered as DETM(+15.99)FNASQR. The "Input tips" field at the bottom of the software provides guides with regards to the aforementioned parameters that have to be entered in the "input data" menu. In this study, decoys were generated at a target-to-decoy ratio of 1:20, the "mass/charge

tolerance" was set at 20 ppm, the number of decoys from each category was 3, and iodoacetamide was selected for Cys modification of all Cys-containing target glycopeptides.

Once all the “Input Data” parameters have been entered, and the user clicks the “Generate Decoy Glycopeptides” button, the software generates the decoys. The decoy list is displayed in the “Result” page of the software, and an example output file is shown in Figure 1B. The figure shows 20 glycopeptide decoys generated for the target glycopeptide and input parameters displayed in Figure 1A. The decoys have varying glycan compositions, and their monoisotopic *m/z's* are close (within 20 ppm) to that of the target glycopeptide.

3.3.2 How GDG Generates Glycopeptide Decoys

Figure 2 shows a schematic representation of the approach used by GDG to generate decoys. To create a decoy for any target glycopeptide, GDG uses two main steps. First, a glycan is randomly selected from a library of over 300 biologically relevant *N*-linked glycans, and secondly, a peptide mass is generated. To choose a glycan for the decoy, the software queries a library of glycans that has been parsed into three categories (described in the preceding section), so that decoys of diverse glycan compositions could be easily generated. After a random glycan is selected, the algorithm determines if the selected glycan had been previously picked to generate a decoy for the same target. If so, the software discards the glycan and selects a new glycan. After selecting a non-redundant glycan from the library, the second step of decoy generation is to identify an appropriate peptide mass. The mass that represents the peptide portion of the decoy is computed such that the *m/z* of the entire decoy (peptide + glycan portions) is within the user-specified mass tolerance from the *m/z* of the precursor ion. The randomly selected glycan plus its arbitrary “peptide” mass represents the decoy. One final restriction is

placed on the decoy: The mass appended to the glycan must not be smaller than the sum of the monoisotopic masses of asparagine and lysine amino acid residues. Once a decoy is generated, it is added to a restriction list to avoid duplication, and more decoys are subsequently generated until the user-specified number of decoys has been created. GDG can generate up to 45 decoys per target. The decoys are generated irrespective of the glycan type on the target or the enzyme used for digestion of the glycoprotein.

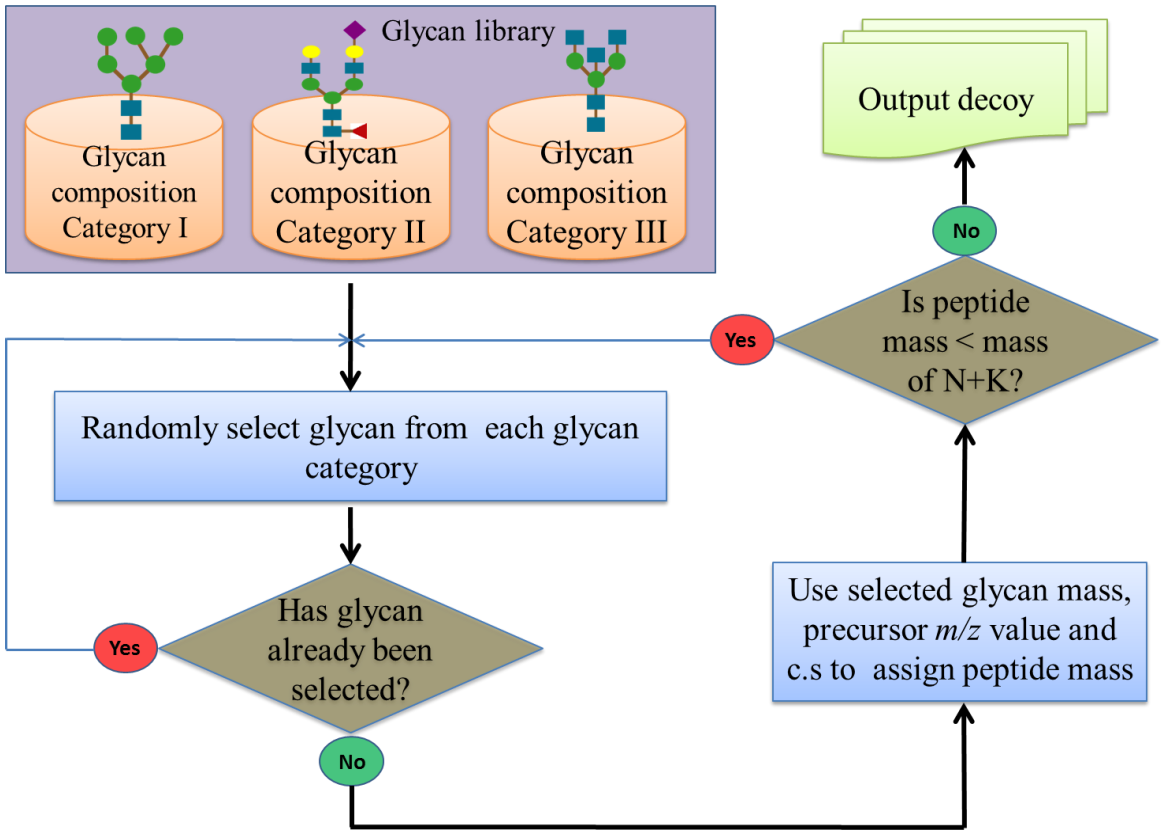


Figure 2. Schematic representation of the decoy generation approach used by GDG. Decoys are generated via two main steps: First, a glycan is randomly selected from a pool of about 300 biologically relevant *N*-linked glycans separated in three categories (see text); and second, an arbitrary mass, representing the decoy's peptide portion, is added to the glycan so that the total mass of the decoy is within a user-specified mass tolerance from the target glycopeptide mass.

3.3.3 Tool 2: The Glycopeptide Spectral Library

We collected a significant dataset of 100 CID spectra of known glycopeptide compositions; these spectra can be used for testing any glycopeptide scoring algorithm that accepts CID spectra. Figure 3 shows one example CID spectrum in the dataset. This CID spectrum is from IgG1 monoclonal antibody, and the data was assigned to the glycopeptide composition EEQYNSTYR+[Hex]3[HexNAc]5[Fuc]1. This glycopeptide had been previously assigned as being present in this particular protein.[33] Along with all spectra in the dataset, the monoisotopic *m/z* of this glycopeptide matches the theoretical mass quite closely. The doubly charged precursor ion, *m/z* 1419.0697, is within 2 ppm from the theoretical value, *m/z* 1419.0663. In addition to matching the high-resolution mass, inspection of the CID data indicates that oxonium ions at *m/z* 528.2 and 690.1 are present, further confirming that the precursor ion is indeed a glycopeptide. By also identifying the Y_1 ion, and confirming neutral loses of monosaccharide residues from the potential glycopeptide, the spectrum was assigned as EEQYNSTYR+[Hex]3[HexNAc]5[Fuc]1. In a similar manner, the 100 CID spectra in the MS/MS dataset were unambiguously assigned to their known glycopeptide compositions. The dataset consists of spectra from 33 high-mannose glycopeptides, 34 fucosylated glycopeptides and 33 non-fucosylated complex/hybrid glycopeptides. The glycopeptide compositions assigned to these spectra are henceforth referred to as target glycopeptides (or targets). The peaklists for all CID spectra, and their assigned glycopeptide compositions, are available at http://pubs.acs.org/doi/suppl/10.1021/acs.jproteome.7b00289 (the Supporting Information of the original publication included in this chapter). In addition, 13 of the spectra are annotated. See Supplemental Figures S1 to S13 from the aforementioned website.

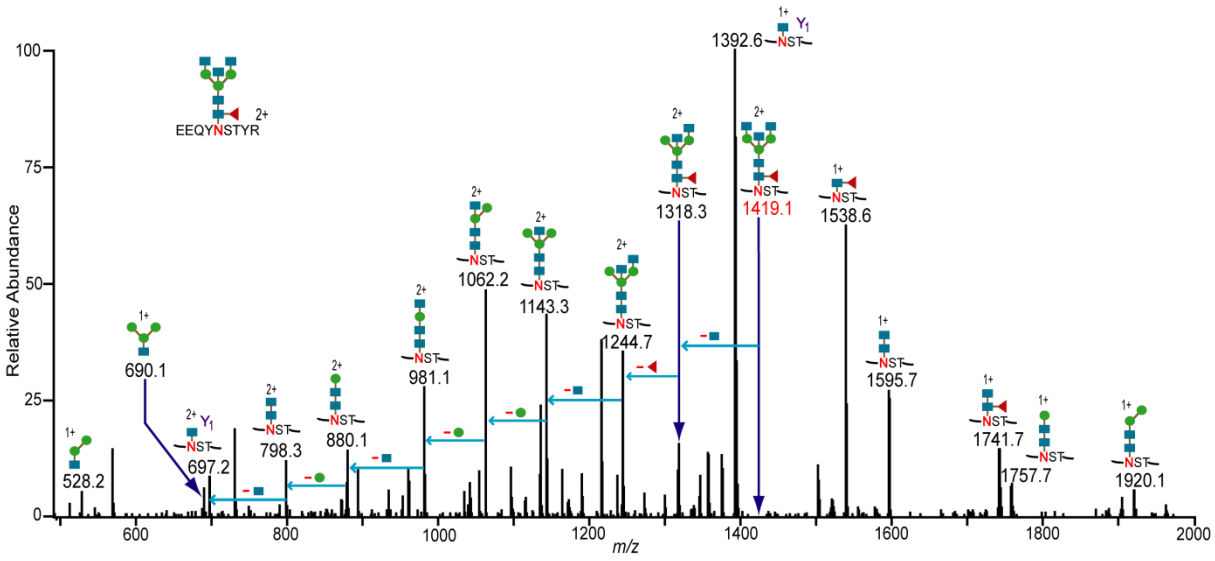


Figure 3. Representative CID spectrum showing the assignment of an IgG1glycopeptide (shown in figure). Glycosidic cleavages of the glycan portion and glycan oxonium ions are observed. The Y_1 product ion and neutral monosaccharide losses from the potential glycopeptide candidate (at *m/z* 1419.1) were used to confirm the assignment. The 100 spectra in the "Glycopeptide Spectral Library" were assigned to their correct glycopeptide compositions in the same way. The blue squares, green circles and red triangles denote *N*-acetylhexoseamine, hexose, and fucose monosaccharide residues, respectively.

3.3.4 Application of Tools 1 and 2 for Evaluating the Accuracy of CID Scoring Algorithms for Glycopeptides

Figure 4 shows how the two new tools described above (Glycopep Decoy Generator and the large dataset of CID spectra) can be used to assess the accuracy of software that assign glycopeptides to CID data.

The dataset of 100 CID spectra described in Section 3.3.3 of this dissertation and known targets for each spectrum was used in conjunction with the Decoy Generator to demonstrate how these tools are helpful in testing and refining glycopeptide analysis software. Specifically, the scoring algorithm of GlycoPep Grader[21] was evaluated herein. GlycoPep Grader uses

monosaccharide neutral loses from glycopeptide compositions to score target and decoy glycopeptides against a CID spectrum, and the candidate with the highest score is assigned to the spectrum in question.[21] For each of the 100 target glycopeptide compositions, and the 100 assigned spectra, 20 decoys were generated using the Decoy Generator, and GlycoPep Grader was used to score each target and its 20 decoys against the known CID spectrum of the target. The scores were interrogated to determine whether GlycoPep Grader consistently matched the CID spectra to their correct glycopeptide compositions or to decoy glycopeptides.

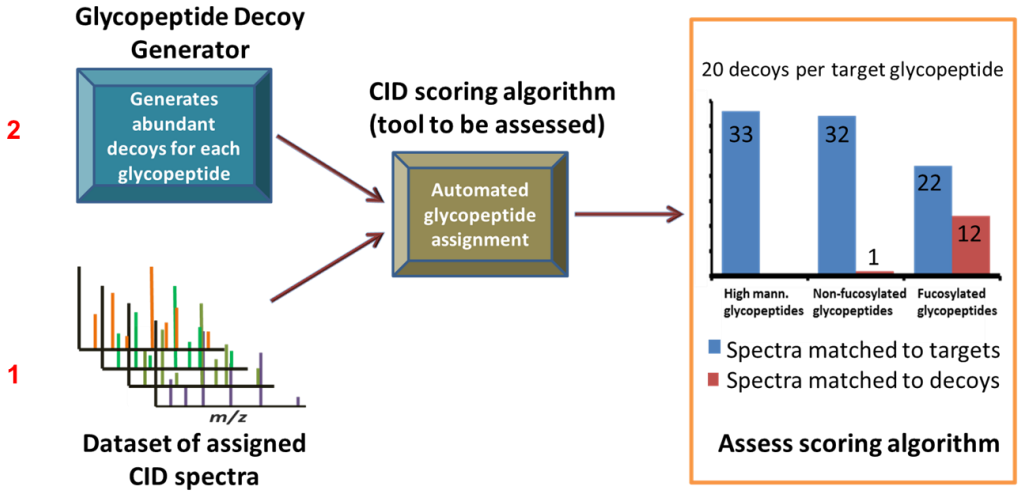


Figure 4. Schematic representation of a simple approach to assess the accuracy of CID scoring algorithms. A tool is tested by using a large dataset of CID spectra **(1)**, abundant decoys for each target glycopeptide in the dataset generated using the Decoy Generator **(2)**. The spectra in the library scored against their known targets along with the abundant decoys generated by GDG. Algorithms that have some weaknesses will match more spectra to decoys.

A summary of the results is shown in Figure 5A. As shown, all the 33 spectra originating from high-mannose glycopeptides were matched to their correct targets; one out of the 33 spectra (3%) from non-fucosylated glycopeptides was matched to a decoy glycopeptide (32 spectra matched

the correct targets), while 12 out of the 34 spectra (35%) from fucosylated glycopeptides were matched to decoys (22 spectra matched to the correct targets). The results clearly indicate that GlycoPep Grader accurately scores spectra of high-mannose and non-fucosylated complex/hybrid glycopeptide compositions, but it has some weaknesses in scoring spectra of fucosylated glycopeptides. Hence, GlycoPep Grader's scoring algorithm can be modified to improve its accuracy in scoring spectra of fucosylated glycopeptide compositions. The annotated CID spectra of the 13 glycopeptides that were incorrectly assigned to decoys are available at http://pubs.acs.org/doi/suppl/10.1021/acs.jproteome.7b00289 (the Supporting Information of the original publication included in this chapter).

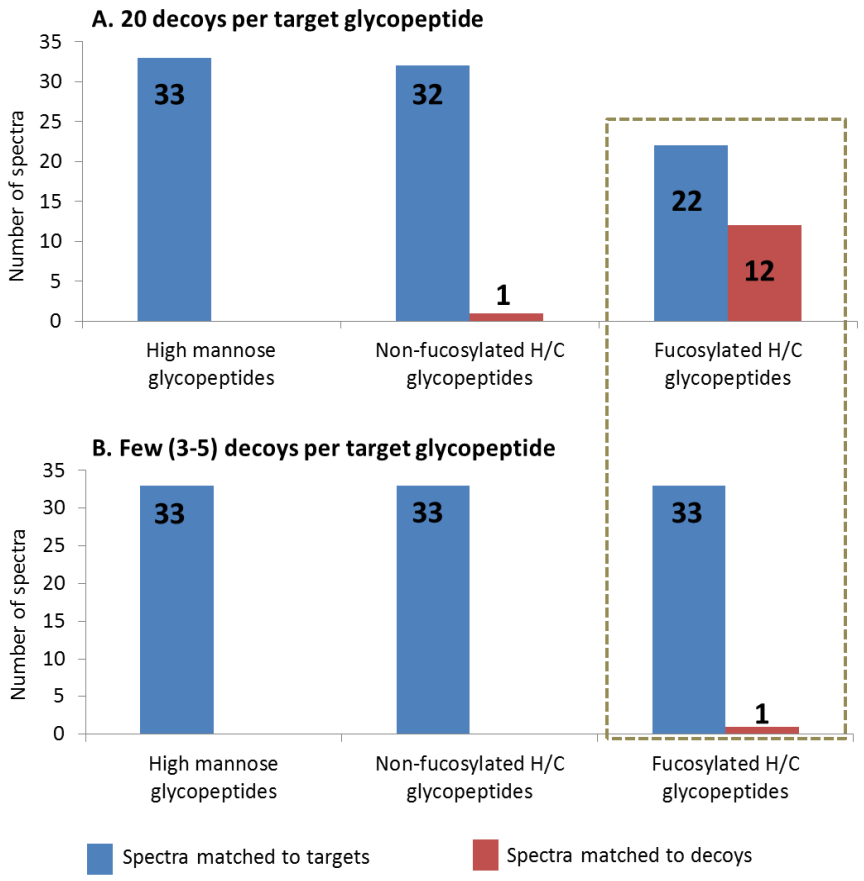


Figure 5. Bar graphs showing results of scoring 33 high mannose, 33 non-fucosylated hybrid/complex (H/C), and 34 fucosylated hybrid/complex glycopeptide spectra scored against the known target glycopeptide compositions along with **(A.)** abundant decoys (20 decoys) per target glycopeptide and **(B.)** few decoys (three to five) decoys per target glycopeptide. More spectra of fucosylated glycopeptides were matched to decoys when abundant decoys were used (box).

3.3.5 Are Abundant Decoys Needed for Accurate Evaluation of Glycopeptide Scoring Algorithms?

After using abundant decoys, generated by the Decoy Generator, to identify the weakness in GlycoPep Grader's scoring algorithm, we wanted to determine why the limitation was not identified during the development of the software. In the original publication describing GlycoPep Grader,[21] a total of 79 glycopeptides were scored using the software, and all 79

glycopeptides, including 17 fucosylated glycopeptides, were correctly assigned to the known glycopeptide spectra, even when scoring each spectrum against decoys. In that work, typically 3-5 decoys were used. Hence, for this case study, we replicated the procedure used to generate decoys during the initial development of GlycoPep Grader, and we scored those decoys and targets against our new dataset of 100 CID spectra. We used the same number of decoys per target glycopeptide (three to five decoys), and we generated the decoys in the same manner as described previously: For each of the 100 target glycopeptide compositions, either three, four, or five decoys were generated from Titin, a glycoprotein containing about 50,000 amino acid residues, and a database of about 200 glycans that were multiplexed to the protein *in silico,* as described by Woodin *et al*.[21] Each target and its decoys were scored against the known CID spectrum of the target using GlycoPep Grader. The results are shown in Figure 5B. The 33 spectra of high-mannose glycopeptides and the 33 spectra of non-fucosylated glycopeptides were all matched to their correct glycopeptide compositions, and only one out of the 34 spectra of fucosylated glycopeptides was matched to a decoy. Hence, of the 100 target glycopeptides scored with limited numbers of decoys, 99 were matched to the correct target glycopeptides and only one spectrum was matched to a decoy. This result is contrary to the aforementioned case when the target-to-decoy ratio was 1:20, and up to 13 spectra (12 of which were of fucosylated glycopeptides) were assigned to decoys. A comparison of the scores of the 12 spectra from fucosylated glycopeptides that were assigned to decoys when scored at a target-to-decoy ratio of 1:20 and the scores of the same spectra when scored against fewer decoys is provided in Supplemental Tables 1-12 of the original publication that is included in this chapter (available at http://pubs.acs.org/doi/suppl/10.1021/acs.jproteome.7b00289). Table 1 shows an example of the

scores of one of the 12 spectra that was assigned to the target glycopeptide when scored against a few decoys (Table 1A) but it was wrongly assigned to a decoy when scored at a target-to-decoy ratio of 1:20 (Table 1B).

Overall, the results indicate that GlycoPep Grader's limitation in scoring spectra from fucosylated glycopeptides could not be determined by scoring target glycopeptide spectra against a limited number of decoys, which explains why GlycoPep Grader's weakness was not identified during the development of the tool, when between three to five decoys per target were used. Hence, abundant decoys are indeed needed for accurate assessment of tools that assign glycopeptides to MS/MS spectra.

Given the above results, it is imperative for software developers to use large numbers of decoys to test their scoring algorithms during the development of software designed to match glycopeptides to MS/MS data. By so doing, the probability of decoy matches increases, and when decoys outscore glycopeptide candidates that are known to be correct, software developers can easily make changes to improve the scoring algorithm. Similarly, end-users of glycopeptide software can assess the quality of the output from various tools by testing them against large spectral libraries of known glycopeptide compositions and large numbers of decoys per target glycopeptide. However, generating a large spectral library is time consuming. Hence, the peak lists of all 100 spectra used in this study have been provided in the supplementary information.

Table 1: Results of scoring a fucosylated target glycopeptide against its known spectrum (shown in Figure 3) together with **(A.)** 5 decoys and **(B.)** 20 decoys

A. Target-to-decoy ratio of 1:5

Name	Composition	Mono *m/z*	Score
Target	EEQYNSTYR+[Hex]3[HexNAc]5[Fuc]1	1419.0663	83.08
Decoy1	AYANVSSKCSK +[Hex]5[HexNAc]4	1419.0862	20.22
Decoy2	VNVSSSK+[Hex]5[HexNAc]5[NeuNAc]1	1419.0762	16.50
Decoy3	YQSNATLVCK+[Hex]5[HexNAc]3[NeuNAc]1	1419.0805	16.50
Decoy4	NTSDVMYKK+[Hex]4[HexNAc]4[NeuNAc]1	1419.0805	16.50
Decoy5	QNQTLYSQK+[Hex]6[HexNAc]3[Fuc]1	1419.0894	0.00

B. Target-to-decoy ratio of 1:20

Name	Composition	Mono *m/z*	Score
Target	EEQYNSTYR+[Hex]3[HexNAc]5[Fuc]1	1419.0663	83.08
Decoy1	**Peptide (1334.5734)+[Hex]3[HexNAc]5**	**1419.0717**	**83.25**
Decoy2	Peptide (1131.4316)+[Hex]3[HexNAc]6	1419.0405	40.44
Decoy3	Peptide (485.2852)+[Hex]5[HexNAc]4[NeuNAc]2[Fuc]1	1419.065	33.50
Decoy4	Peptide (661.2745)+[Hex]5[HexNAc]6[Fuc]1	1419.0437	29.37
Decoy5	Peptide (834.3829)+[Hex]5[HexNAc]3[NeuNAc]2	1419.0453	28.68
Decoy6	Peptide (938.4758)+[Hex]4[HexNAc]4[NeuNAc]1[Fuc]1	1419.0862	26.80
Decoy7	Peptide (1594.6456)+[Hex]3[HexNAc]3[Fuc]1	1419.0573	23.93
Decoy8	Peptide (1959.7656)+[Hex]2[HexNAc]2[Fuc]1	1419.0513	22.33
Decoy9	Peptide (1133.5523)+[Hex]8[HexNAc]2	1419.0741	18.21
Decoy10	Peptide (280.2338)+[Hex]7[HexNAc]7	1419.0869	16.50
Decoy11	Peptide (1416.5875)+[Hex]5[HexNAc]3	1419.0521	16.50
Decoy12	Peptide (340.2967)+[Hex]5[HexNAc]4[NeuNAc]3	1419.0895	16.50
Decoy13	Peptide (266.1611)+[Hex]6[HexNAc]5[NeuNAc]2	1419.0402	16.50
Decoy14	Peptide (395.29485)+[Hex]7[HexNAc]5[NeuNAc]1	1419.0857	16.50
Decoy15	Peptide (1943.842)+[Hex]3[HexNAc]2	1419.0869	11.17
Decoy16	Peptide (499.2631)+[Hex]6[HexNAc]6[Fuc]1	1419.0644	10.05
Decoy17	Peptide (1781.7508)+[Hex]4[HexNAc]2	1419.0677	8.38
Decoy18	Peptide (1797.7531)+[Hex]3[HexNAc]2[Fuc]1	1419.0714	8.38
Decoy19	Peptide (1619.6457)+[Hex]5[HexNAc]2	1419.0415	0.00
Decoy20	Peptide (1457.6558)+ [Hex]6[HexNAc]2	1419.073	0.00

Note: The annotated CID spectrum of the target glycopeptide is shown in Figure 3. The decoy that outscores the target glycopeptide is shown in blue.

3.4 Conclusion

To simplify the task of building effective glycopeptide software, we developed two new tools, Glycopeptide Decoy Generator (GDG) and an expert-assigned dataset of 100 CID spectra. GDG rapidly generates glycopeptide decoys *de novo,* and these decoys can be used to assess the quality of tools that assign glycopeptides to CID data. As a secondary contribution, we provide herein peak lists for 100 validated CID spectra that can be used to test any existing software tool or any new tool under development. Using large numbers of decoys generated by our newly developed tool, and our set of 100 validated CID spectra, we evaluated the accuracy of existing software that assigns glycopeptides to CID data. We demonstrate that limitations in the scoring algorithm of the software can be identified when testing it against large sets of decoys, and these limitations could not be identified when only a few decoys were scored.

Our tool is the first software that automatically generates abundant decoys on demand for the assessment of algorithms that assign glycopeptides to CID spectra. The approach for decoy generation is simple; it can be used as-is, or the software can be easily incorporated into other bioinformatics tools designed to match glycopeptides to CID data. The software can be downloaded at http://glycopro.chem.ku.edu/GPJ.jar.

3.5 Acknowledgement

The authors acknowledge funding from the National Institutes of Health, grant R01GM103547 to Heather Desaire.

3.6 References

1. Boscher, C.; James, W. D.; Nabi, I. R. *Curr. Opin. Cell Biol.*, **2011**, 23, 383-392.
2. Haines. N.; Irvine, K. D. *Nat. Rev. Mol. Cell Biol.*, **2003**, 4, 786-797.

3. Gallagher, P.; Henneberry, J.; Wilson, I.; Sambrook, J.; Gething, M. J. *J. Cell. Biol.*, **1988**, 107, 2059-2073.
4. Xu, C.; Ng, D. T. *Nat. Rev. Mol. Cell. Biol.*, **2015**, 16, 742-752
5. Helenius, A.; Aebi, M. *Science.*, **2001**, 291, 2364-2369.
6. Yan, L.;Yuan, T.;Taha, R.; Prakash, A.; Lopez, M. F.; Chan, D. W. Zhang, H. *Anal. Chem.*, **2011,** 83, 240-245.
7. Peracaula, R.; Tabarés, G.; Royle, L.; Harvey, D. J.; Dwek, R. A.; Rudd, P. M.; De Llorens, R. *Glycobiology*, **2003**, 13, 457-470.
8. Jefferis, R. *Trends Pharmacol. Sci.*, **2009**, 30, 356-362.
9. Liu, L. *J. Pharm. Sci.*, **2015**, 104, 1866-1884.
10. Chung, C. H.; Mirakhur, B.; Chan, E.; Le, Q. T.; Berlin, J.; Morse, M.; Murphy, B. A.; Satinover, S. M.; Hosen, J.; Mauro, D.; Slebos, R. J.; Zhou, Q.; Gold, D.; Hatley, T.; Hicklin, D. J.; Platts-Mills, T. *N. Engl. J. Med.*, **2008**, 358, 1109-1117.
11. Solá, R. J.; Griebenow, K. *J. Pharm. Sci.*, **2009**, 98, 1223-1245.
12. Leymarie, N.; Zaia, J. *Anal Chem.*, **2012**, 84, 3040-3048
13. Morelle, W.; Michalski, J. C. *Nat. Protoc.*, **2007**, 2, 1585-1602.
14. Aich, U.; Lakbub, J.; Liu, A. *Electrophoresis.*, **2016**, 37, 1468-1488.
15. Dalpathado, D. S.; Desaire, H. *Analyst*, **2008** 133, 731-738.
16. Zhu, Z.; Desaire, H. *Annu. Rev. Anal. Chem.*, **2015**, 8, 463-483.
17. Chen-Chun, C.; Su, Wan-Chih.; Huang, Bao-Yu.; Chen, Yu-Ju.; Tai, Hwan-Ching.; Obena, R. *Analyst*, **2014**, 139, 688-704.
18. Bodnar, E.; Perreault, H. *Anal. Chem.*, **2013**, 85, 10895-10903.

19. Froehlich, J. W.; Dodds, E. D.; Wilhelm, M.; Serang, O.; Steen, J. A.; Lee, R. S. *Mol. Cell Proteomics,* **2013**, 12, 1017-1025.

20. Hu, W.; Su, X.; Zhu, Z.; Go, E. P.; Desaire, H. *Anal. Bioanal. Chem.,* **2017**, 409, 561-570.

21. Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. *Anal. Chem.,* **2012**, 84, 4821-4829.

22. Chandler, K. B.; Pompach, P.; Goldman, R.; Edwards, N. *J. Proteome Res.,* **2013**, 12, 3652-3666.

23. Lynn, K. S.; Chen, C. C.; Lih, T. M.; Cheng, C. W.; Su, W. C.; Chang, C. H.; Cheng, C. Y.; Hsu, W. L.; Chen, Y. J.; Sung, T. Y. *Anal Chem.*, **2015**, 87, 2466-2473.

24. Toghi, E. S.; Shah, P.; Yang, W.; Li, X.; Zhang, H. *Anal Chem.,* **2015**, 87, 5181-5188.

25. Zeng, W. F.; Liu, M. Q.; Zhang, Y.; Wu, J. Q.; Fang, P.; Peng, C.; Nie, A.; Yan, G.; Cao, W.; Liu, C.; Chi, H.; Sun, R. X.; Wong, C. C.; He, S. M.; Yang, P. *Sci. Rep.*, **2016,** 6, 25102.

26. Zhu, Z.; Hua, D.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.,* **2013**, 85, 5023-5032.

27. Zhu, Z.; Su, X.; Go, E. P.; Desaire, H. *Anal. Chem.,* **2014**, 86, 9212-9219.

28. Mayampurath, A.; Yu, C. Y.; Song, E.; Balan, J.; Mechref, Y.; Tang, H. *Anal. Chem.*, **2014**, 86, 453-463.

29. He, L.; Xin, L.; Shan, B.; Lajoie, G. A.; Ma, B. *J. Proteome Res.*, **2014**, 13, 3881-3895.

30. Hu, H.; Khatri, K.; Zaia, J.; *Mass Spectrom. Rev.,* **2017,** 36, 475-498.

31. Woodin, C. L.; Maxon, M.; Desaire, H. *Analyst,* **2013**, 138, 2793-2803.

32. Hu, H.; Khatri, K.; Klein, J.; Leymarie, N.; Zaia, J. *Glycoconj. J.,* **2016**, 33, 285-296.

33. Wuhrer, M.; Stam, J. C.; Van de Geijn, F. E.; Koeleman, C. A.; Verrips, C. T.; Dolhain, R. J.; Hokke, C. H.; Deelder, A. M. *Proteomics,* **2007**, 7, 4070-4081.

34. Alley, W. R.; Mechref, Y.; Novotny, M. V. *Rapid Commun. Mass Spectrom.*, **2009**, 2, 161-170.

35. Brown, K. J.; Vanderver, A.; Hoffman, E. P.; Schiffmann, R.; Hathout, Y. *Int. J. Mass Spectrom*., **2012**, 312, 97-106.

36. Go, E. P.; Chang, Q.; Liao, H. X.; Sutherland, L. L.; Alam, S. M.; Haynes, B. F.; Desaire, H. *J. Proteome Res*., **2009,** 8, 4231-4242.

37. Liu, X.; McNally, D. J.; Nothaft, H.; Szymanski, C. M.; Brisson, J. R.; Li, J. *Anal. Chem*., **2006**, 78, 6081-6087.

38. Go, E. P.; Herschhorn, A.;, Gu, C.; Castillo-Menendez, L.; Zhang, S.; Mao, Y.; Chen, H.; Ding, H.; Wakefield, J. K.; Hua, D.; Liao, H. X.; Kappes, J. C.; Sodroski, J.; Desaire, H. *J. Virol.,* 2015, 89, 8245-8257.

39. Domon, B.; Costello, C. E. *Glyucoconjugate J.,* 1988, 5, 397-409.

40. Jiang, H.; Desaire, H.; Butnev, V. Y.; Bousfield, G. R. *J. Am. Soc. Mass Spectrom*., 2004, 15, 750-758.

41. Go, E. P.; Rebecchi, K. R.; Dalpathado, D. S.; Bandu, M. L.; Zhang, Y.; Desaire, H. *Anal. Chem.,* 2007, 79, 1708-1713.

Chapter 4: CID Fragmentation Patterns of Fucosylated Glycopeptides: Toward an Improved GlycoPep Grader Software

Abstract

GlycoPep Grader is one of many bioinformatics tools developed in recent years to facilitate the tedious task of assigning glycopeptides to MS/MS spectra. GlycoPep Grader assigns glycopeptides to CID data by searching for two types of product ions: (1) [peptide + core component] product ions, which are used to score ions that contain the peptide plus components of the pentasaccharide core, and (2) [precursor – monosaccharide] ions, which are used to score ions resulting from neutral loss of monosaccharide residues from the candidate precursor ion. In Chapter 3, we assessed the accuracy of GlycoPep Grader's scoring algorithm and found that while the tool accurately scores CID spectra from high mannose and non-fucosylated glycopeptides, it had some limitations in scoring spectra from fucosylated glycopeptides. Herein, we study the CID fragmentation characteristics of fucosylated glycopeptides and GlycoPep Grader's scoring rules in an effort to solve the problem of wrongly assigned data from fucosylated glycopeptides. We identified some prominent product ions from a fragmentation pathway that was not considered when writing GlycoPep Grader's rules for scoring fucosylated glycopeptides. Based on this finding, we propose new scoring rules for fucosylated glycopeptide compositions that can be incorporated into GlycoPep Grader to increase the scores of fucosylated glycopeptides. The approach used here to improve the scoring algorithm of GlycoPep Grader

could henceforth be used to improve any other algorithm that assigns glycopeptides based on their MS/MS data.

4.1 Introduction

Glycopeptide analysis is an important approach for protein glycosylation profiling because it provides glycosylation site-specific information.[1,2] However, glycopeptide analysis using MS/MS data is challenging, even after the recent emergence of numerous automated tools that facilitate the process. Several bioinformatics tools have been developed to interpret glycopeptide MS/MS data generated by different fragmentation techniques, including electron transfer dissociation (ETD), higher-energy collisional dissociation (HCD), and collision induced dissociation (CID), the most common fragmentation technique used for glycopeptide analysis.[3-5] Examples of tools that analyze CID glycopeptide data include MAGIC,[6] GlycoPeptideSearch,[7] and GlycoPep Grader;[8] the latter being the focus of this study. Although these tools greatly enhance data analysis speed and reduce turn-around time for protein glycosylation profiling, automated glycopeptide analysis could have some drawbacks. For example, an assessment of the accuracy of GlycoPep Grader in Chapter 3 of this dissertation revealed that it has some limitation in scoring CID spectra from fucosylated glycopeptides, and this weakness could be detrimental in estimating fucosylation in glycoproteins.

Accurate determination of the fucosylation levels of glycoproteins is vital in understanding some critical biological activities of glycoproteins because the degree of fucosylation can affect the bioactivity of some glycoproteins. For example, an increase in antibody fucosylation can lead to decreased antibody dependent cellular cytotoxicity (ADCC).[9-11] Glycopeptides can be fucosylated at two main sites on the glycan components: (1) core

fucosylation, where a fucose residue is appended to a GlcNAc residue on the pentasaccharide core of a glycan moiety; and (2) antenna fucosylation where the fucose residue is attached to a GlcNAc residue on one of the glycan's antennae.[12] While antenna fucosylation is rare, core fucosylation is very common, and it is the focus of this study. Although core- and antenna-fucosylated glycopeptide isomers can be distinguished using MS/MS data,[12] a thorough analysis of the fragmentation patterns of fucosylated glycopeptides is still lacking in the literature, making it difficult to write accurate rules for scoring CID spectra from fucosylated glycopeptides.

The development of efficient automated glycopeptide scoring algorithms depends on the comprehensive analysis and identification of the characteristic fragmentation patterns of glycopeptides containing varying glycan components. Hence, writing accurate glycopeptide scoring algorithms is, perhaps, the most challenging part of software development due to the micro heterogeneity of *N*-linked glycans, since the scoring rules must be specific for each glycan type, and they must be based on the predominant product ions of glycopeptide type upon fragmentation.[8] Currently, very limited information is available on the fragmentation behavior of fucosylated glycopeptides because fucose is generally considered to be a very labile monosaccharide.[13] This limitation in the understanding of the characteristic fragmentation behavior of fucosylated glycopeptides could translate into weaknesses in glycopeptide tools, as is the case with GlycoPep Grader. Hence, a thorough study of the characteristic behavior of fucosylated glycopeptides upon CID fragmentation can uncover prominent product ions that could be used for accurate assignment of fucosylated glycopeptides to MS/MS data.

Herein, we use a large dataset containing CID spectra of glycopeptides from well characterized glycoproteins to study the fragmentation characteristics of fucosylated glycopeptides. The overall goal of this study is to identify the source of GlycoPep Grader's limitation in scoring fucosylated glycopeptides, and to suggest possible changes to improve the software. The results of our study show that although the fucose residue is labile, the core fucose does not always dissociate from the glycan during CID fragmentation. In addition to the fragmentation pathway where the fucose residue dissociates from the precursor glycopeptide ion, we identified an additional fragmentation pathway, where the fucose residue remains attached to the glycan core. A study of the current GlycoPep Grader scoring rules indicate that the software does not take this additional fragmentation pathway into consideration when scoring fucosylated glycopeptides. Based on this finding, we propose new scoring rules for fucosylated glycopeptides which could be incorporated into GlycoPep Grader, and any other algorithm that scores CID spectra of glycopeptides, in order to address its limitations reported in Chapter 3 of this dissertation.

4.2 Experimental

The CID data used in this study are the same data used in Chapter 3 of this dissertation. Hence, details regarding the materials and reagents, sample preparation, LC-MS/MS analysis, and generation of the CID dataset are found in the experimental section (Section 3.2) of Chapter 3.

4.3 Results and Discussion

4.3.1 Overview of GlycoPep Grader's Scoring Algorithm

In order to identify and correct GlycoPep Grader's weakness in scoring fucosylated glycopeptides, it is important to understand the overall scoring rules of the software. The detailed scoring algorithm has been reported by Woodin et al.[8] A brief summary of the scoring algorithm is discussed herein. GlycoPep Grader scores potential glycopeptide candidate compositions (including decoy glycopeptides) by searching predicted *in-silico* product ions of the glycopeptide compositions against the experimental CID data. The candidate that best matches the experimental spectrum (candidate with highest score) is assigned to the CID spectrum being studied. The software scores candidate glycopeptide compositions using two predominant types of product ions that have been shown to be present in CID spectra of glycopeptides: the peptide plus pentasaccharide core ions, referred to as [peptide + core component] product ions; and product ions resulting from step-wise neutral losses of terminal monosaccharide residues from the precursor glycopeptide composition, referred to as [precursor – monosaccharide] product ions. The numbers of predicted peptide and glycan ions matched to the experimental spectra are used to calculate the *PeptideScore* and the *GlycanScore*, respectively, and both scores are used to calculate the overall score (*GlycopeptideScore*) of the glycopeptide candidate. The [peptide + core component] and [precursor – monosaccharide] product ions that are searched by the current version of GlycoPep Grader when scoring glycopeptide compositions are shown in Figure 1.

To compute the *PeptideScore*, the software searches for six possible product ions at different charge states. These include: (1) the [naked peptide], (2) [peptide + HexNAc], (3) [peptide + 2HexNAc], (4) [peptide + 2HexNAc + Hex], (5) [peptide + 2HexNAc + 2Hex], and

(6) [peptide + 2HexNAc + 3Hex]. For the *GlycanScore*, the [precursor – monosaccharide] ions searched by GlycoPep Grader depend on the type of glycan component of the glycopeptide candidate. The glycan types are divided into 8 categories: (1) Non-fucosylated high mannose type glycans; (2) fucosylated high mannose type glycans; (3) Non-fucosylated complex or hybrid type glycans containing sialic acid; (4) fucosylated complex or hybrid type glycans containing sialic acid; (5) fucosylated complex or hybrid type glycans that do not contain sialic acid and have more HexNAc than Hex residues; (6) fucosylated complex or hybrid type glycans that do not contain sialic acid and have more Hex than HexNAc residues; (7) complex or hybrid type glycan structures with a greater number of HexNAc than Hex residues and do not contain fucose or sialic acid residues; and (8) complex or hybrid type glycans that have more Hex than HexNAc residues and do not contain fucose or sialic acid residues.

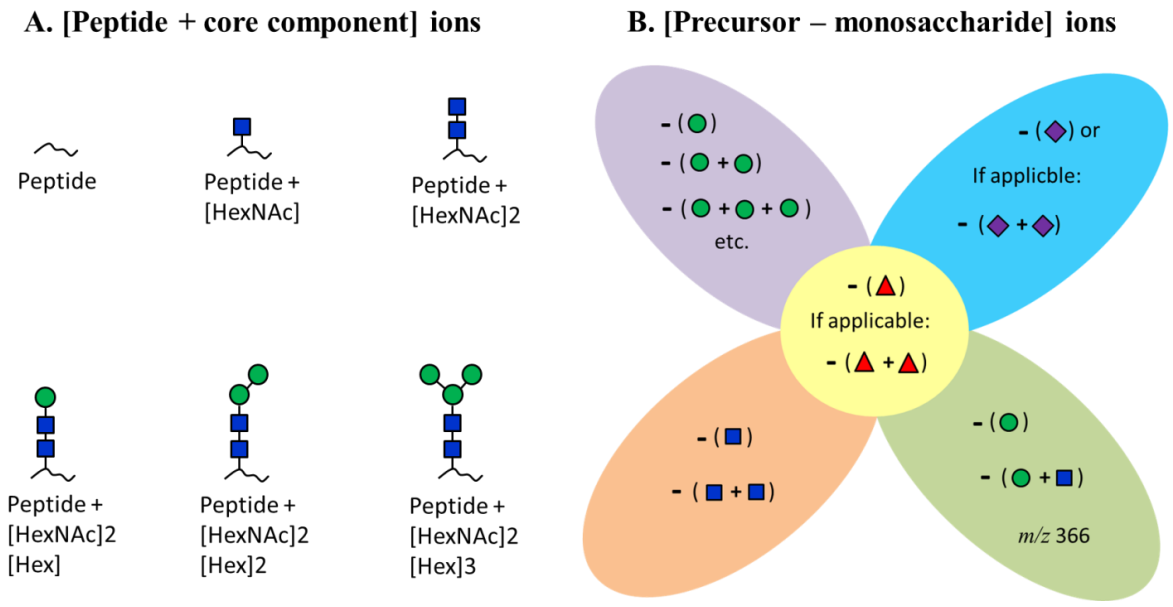


Figure 1. Representation of the predominant product ions searched by GlycoPep Grader when scoring *N*-linked glycopeptide compositions against CID spectra. Six possible [peptide + core component] ions (**A**) are searched for as well as several [precursor-monosaccharide] ions (**B**) for each of the eight groups (see text) of glycopeptides. The monosaccharide neutral losses evaluated for group 1 (see text) are shown in the purple oval. For group 2, the relevant losses are shown in both the purple oval and the yellow circle. For group 3, the relevant losses are shown in the blue oval. For group 4, they are shown in the blue oval and yellow circle. For group 5, they are shown in the orange oval and yellow circle. For group 6, they are shown in the green oval and yellow circle. For groups 7 and 8, neutral losses are in the orange oval at bottom left, and by the green oval at bottom right, respectively. This figure is reprinted with permission from Reference 8.

The rest of this chapter describes how the scoring rules summarized above and the analysis of the fragmentation patterns of CID spectra were used to identify the source of GlycoPep Grader's limitation in scoring spectra from fucosylated glycopeptides.

4.3.2 Evaluation of the Peptide and Glycan Scores of High-mannose, Non-fucosylated, and Fucosylated Glycopeptide Compositions

Since the *GlycopeptideScore* of any glycopeptide candidate is based on the peptide and glycan scores of the glycopeptide composition, we studied the *PeptideScores* and the *GlycanScores* for the 100 target glycopeptide compositions that were generated and scored against their known CID spectra in Chapter 3 of this dissertation. The goal of this particular study was to find out whether GlycoPep Grader's weakness in scoring fucosylated glycopeptides results from the peptide or glycan scoring rules, or both. The glycan and peptide scores can range from zero to one; hence, for each of the three glycopeptide categories (high mannose, non-fucosylated hybrid/complex, and fucosylated hybrid/complex glycopeptide compositions), the peptide and glycan scores were divided into three scoring groups: (1) glycopeptides with a peptide or glycan score of zero, (2) glycopeptides with peptide or glycan scores greater than zero but less than 0.5, and (3) candidates with scores between 0.5 and one. For each of the three glycopeptide categories, the number of glycopeptide compositions with scores in each of the three scoring groups were expressed as a percentage of the total number of glycopeptide candidates in that category. Note, while the focus of this study is on the fucosylated glycopeptides, the scores of both fucosylated and non-fucosylated glycopeptides were interrogated, so that we could compare the scores for the fucosylated species to the scores of glycopeptides that are typically scored correctly, even against 20 decoys.

The results of the analyses of the peptide and glycan scores of the 100 spectra are summarized in Figure 2. As shown in Figure 2A, 100 % (33 out of 33) of the high mannose glycopeptides, 64% (21 out of 33) of the non-fucosylated glycopeptides, and 97% (33 out of 34)

of fucosylated glycopeptides had *PeptideScores between* 0.5 and 1. None of the glycopeptide compositions had a *PeptideScore* of zero. However, 36% of non-fucosylated H/C glycopeptides had a score greater than zero but less than 0.5, while only 3% of fucosylated glycopeptides had a *PeptideScore* in the same range.

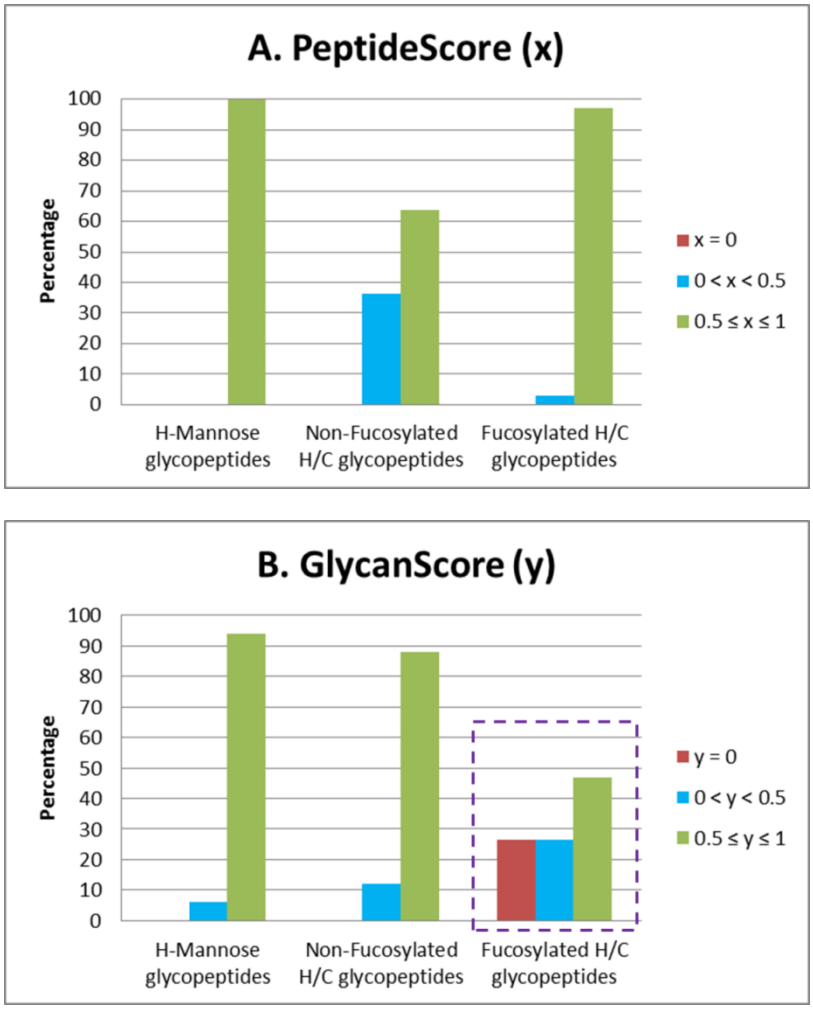


Figure 2. Bar graphs showing results of the evaluation of the peptide (A) and glycan (B) scores of high-mannose, non-fucosylated hybrid/complex (H/C), and fucosylated hybrid/complex type glycopeptide compositions. A high percentage of high-mannose and non-fucosylated hybrid/complex type glycopeptides had peptide and glycan scores between 0.5 and 1; and none of them had a peptide or glycan score of zero. On the contrary, a high percentage (97%) of fucosylated hybrid/complex type glycopeptides

had *PeptideScores* between 0.5 and 1 but up to 26.5% had glycan scores of zero (purple box) and only 47% had glycan scores between 0.5 and 1(purple box).

For the *GlycanScores*, (Figure 2B), 94 % of the high mannose glycopeptides and 88% of the non-fucosylated glycopeptides had a glycan score between 0.5 and 1; while only 47% of fucosylated glycopeptides had a *GlycanScores* between in the same range. Neither the high mannose nor non-fucosylated H/C glycopeptides had a *GlycanScore* of zero. However, up to 27% of the fucosylated glycopeptides had a *GlycanScore* of zero and 27% had *GlycanScores* between 0 and 0.5.

These results were used to evaluate GlycoPep Grader's peptide and glycan scoring rules. A high number of high-mannose and non-fucosylated hybrid/complex glycopeptide compositions that had peptide and glycan scores greater than 0.5; these species are typically correctly assigned. By comparison, many (97%) of the fucosylated hybrid/complex glycopeptide compositions had *PeptideScores* of 0.5 or higher, but up to 26% had a *GlycanScore* of zero, and a total of 53% had *GlycanScores* of less than 0.5, indicating the peptide scoring rules for fucosylated glycopeptides are well written, while there are some short-comings with the glycan scoring rules. Hence, GlycoPep Grader's weakness in scoring fucosylated glycopeptide composition, as identified in Chapter 3 of this dissertation, can be ameliorated by improving the scoring rules for the [precursor – monosaccharide] product ions from fucosylated glycopeptide compositions.

4.3.3 CID Fragmentation Characteristics of N-linked Glycopeptides

After analyzing the peptide and glycan scores, we studied the CID fragmentation patterns and the scoring rules for high-mannose and non-fucosylated hybrid/complex glycopeptide compositions, which are accurately scored by GlycoPep Grader, and compared them to those of fucosylated glycopeptide compositions. The aim was to verify if the scoring rules correlate to the characteristic product ions of each of the glycopeptide types upon CID fragmentation.

4.3.3.1 Fragmentation Characteristics of High-mannose and Non-fucosylated Complex/hybrid Glycopeptide Compositions

Figure 3 shows example CID spectra that exhibit the characteristic fragmentation patterns of high-mannose and non-fucosylated hybrid/complex type glycopeptide compositions. CID fragmentation of high-mannose type glycopeptides were characterized by neutral losses of Hex residues from the precursor glycopeptide ion (Figure 3A), while the fragmentation of non-fucosylated hybrid/complex type glycopeptides were characterized by neutral losses of Hex and HexNAc residues from the precursor ion (Figure 3B). No alternative fragmentation pathways were identified. Both spectra contain the [peptide + core component] (peaks with black star) and the [precursor – monosaccharide] product ions (peaks with red star) which are used by GlycoPep Grader to score glycopeptide compositions. The [peptide + core component] ions were present in both the charge state of the precursor glycopeptide and at lower charge states, while the [precursor – monosaccharide] product ions were mainly present at the charge state of the precursor ion.

After studying the CID spectra and confirming the characteristic product ions, we investigated whether or not GlycoPep Grader's scoring algorithm is designed to search for the characteristic

product ions in the CID spectra of high-mannose and non-fucosylated hybrid/complex type glycopeptides. When calculating the *PeptideScores* for both types of glycopeptide compositions, GlycoPep Grader searches for the six possible [peptide + core component] product ions at both the charge state of the precursor ion and at decremented charge states.

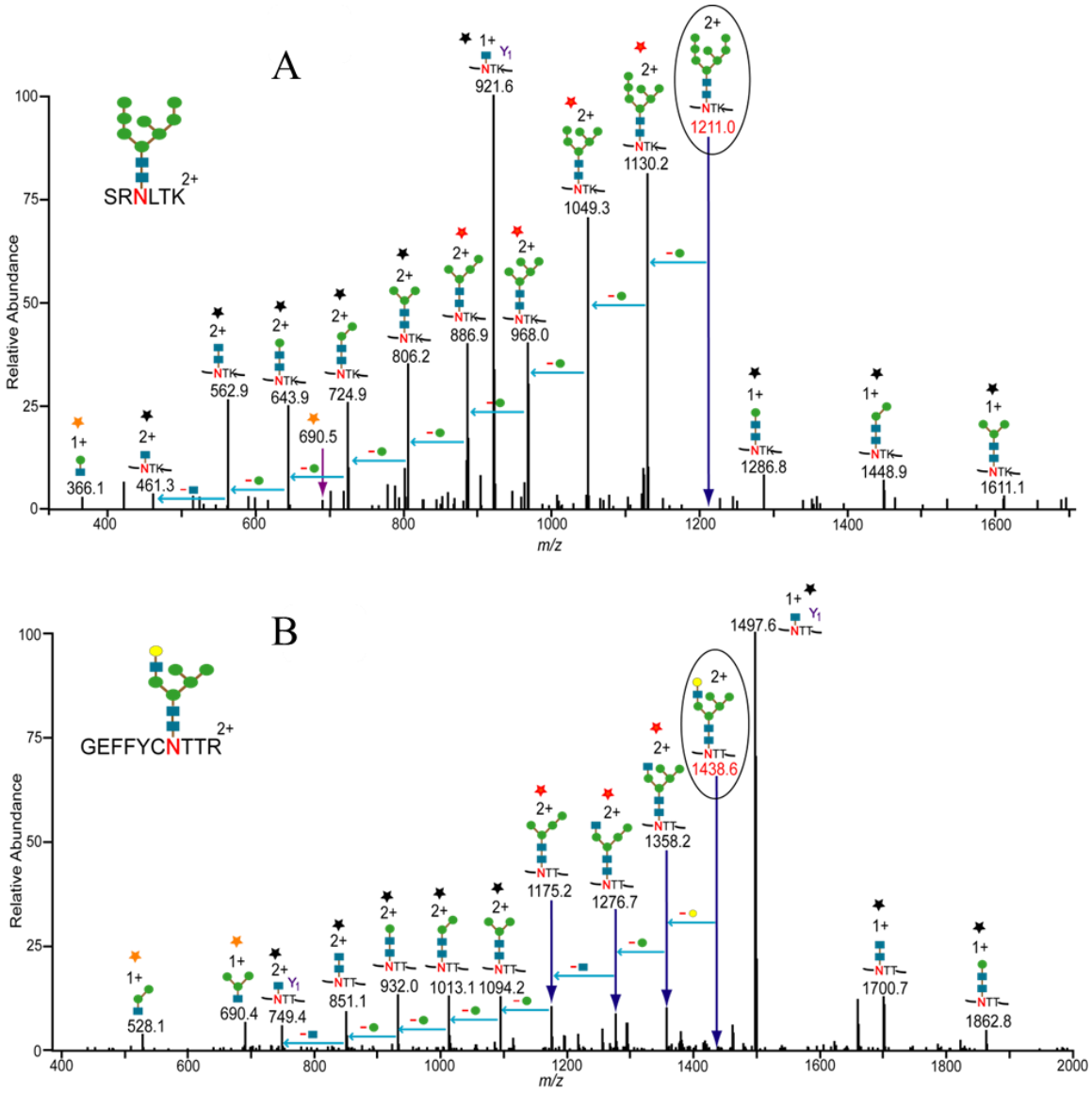


Figure 3. CID spectra of a high-mannose glycopeptide from RNase B (**A**) and a non-fucosylated hybrid type glycopeptide from HIV-1 envelope glycoprotein, C.97ZA012 gp140 (**B**). The spectra show the characteristic fragmentation patterns of high-mannose and non-fucosylated H/C type glycopeptides. High-

mannose spectra (**A**) are characterized by neutral losses of Hex residues from the precursor ion (black oval) and spectra of non-fucosylated H/C type glycopeptides (**B**) are characterized by neutral losses of Hex and HexNAc residues from the precursor ion (black oval). Both spectra contain [peptide + core component] ions (black stars) and [precursor – monosaccharide] product ions (red stars) which are used by GlycoPep Grader to score *N*-linked glycopeptides. Different oxonium ions (orange stars) are also observed, and no alternative fragmentation pathway was identified.

For the *GlycanScores,* the software searches for the predominant production ions resulting from the neutral losses of Hex residues (for high-mannose glycopeptides) and Hex and HexNAc residues (for non-fucosylated glycopeptides) from the glycopeptide precursor ions. Overall, our study of GlycoPep Grader's scoring algorithm for high-mannose and non-fucosylated glycopeptides indicates that the rules are well written, since the software searches for the product ions which are typically present in the CID spectra from these types of glycopeptide compositions. The well-written rules correlate with the high numbers of high-mannose and non-fucosylated glycopeptides that out-scored decoy glycopeptides when GlycoPep Grader's scoring algorithm was assessed in Chapter 3 of this dissertation.

We next studied the CID spectra and GlycoPep Grader's scoring rules for fucosylated glycopeptide compositions to investigate if the same correlation exists between the characteristic product ions in spectra of fucosylated glycopeptides and their scoring rules.

4.3.3.2 Fragmentation Characteristics of Fucosylated Glycopeptides

For a thorough study of the fragmentation patterns of core-fucosylated glycopeptides under CID conditions, two sets of model fucosylated glycopeptides from well characterized proteins were used. These include fucosylated glycopeptides that contain sialic acid and fucosylated glycopeptide compositions that do not contain sialic acid. After extensive analysis of

CID data from these model glycopeptide compositions, we identified two characteristic fragmentation pathways. One pathway results in product ions without appended fucose while the other generates product ions containing the core fucose residue.

A representative CID spectrum of a fucosylated glycopeptide containing sialic acid that exhibits both fragmentation pathways is shown in Figure 4. Figure 4A shows the fragmentation pathway that occurs when the labile sialic acid and fucose residues dissociate from the glycan component. The predominant [precursor-monosaccharide] product ion (e.g peak at *m/z* 1486.4) resulted from the neutral loss of sialic acid from the glycopeptide precursor ion (black oval) and subsequent neutral loss of the fucose residue, followed by neutral losses of Hex and HexNAc residues from the glycan's antennae led to several [precursor – monosaccharide] product ions (peaks indicated with red star) and [peptide + core component] product ions (peaks indicated with black star) that do not contain the fucose residue. The most abundant ion that does not contain fucose residue was the Y1 ion (e.g peak at *m/z* 1364.8) at a charge state lower than that of the precursor ion.

Figure 4B shows the second fragmentation pathway, during which the sialic acid residue dissociates from the glycan component, but the core-fucose residue does not. The predominant [precursor-monosaccharide] product ion (peak at *m/z* 1486.4) also resulted from the loss of sialic acid from the glycopeptide precursor ion (black oval), and the subsequent step-wise neutral losses of monosaccharide residues from the glycan's antennae, without loss of fucose, led to product ions that have the appended core fucose residue. The most abundant ion that contains the appended fucose residue was the Y1+Fuc ion (e.g peak at *m/z* 1510.9) at a charge state lower than that of the precursor ion.

For both fragmentation pathways, the [precursor – monosaccharide] product ions were observed only at the charge state of the precursor glycopeptide while the [peptide + core component] product were observed at both the charge state of the precursor and the next lower charge state. The characteristic oxonium ion for sialylated glycopeptides at *m/z* 657[14,15] was also observed, as well as other common oxonium ions (at *m/z* 528 and 690) of *N*-linked glycopeptides.

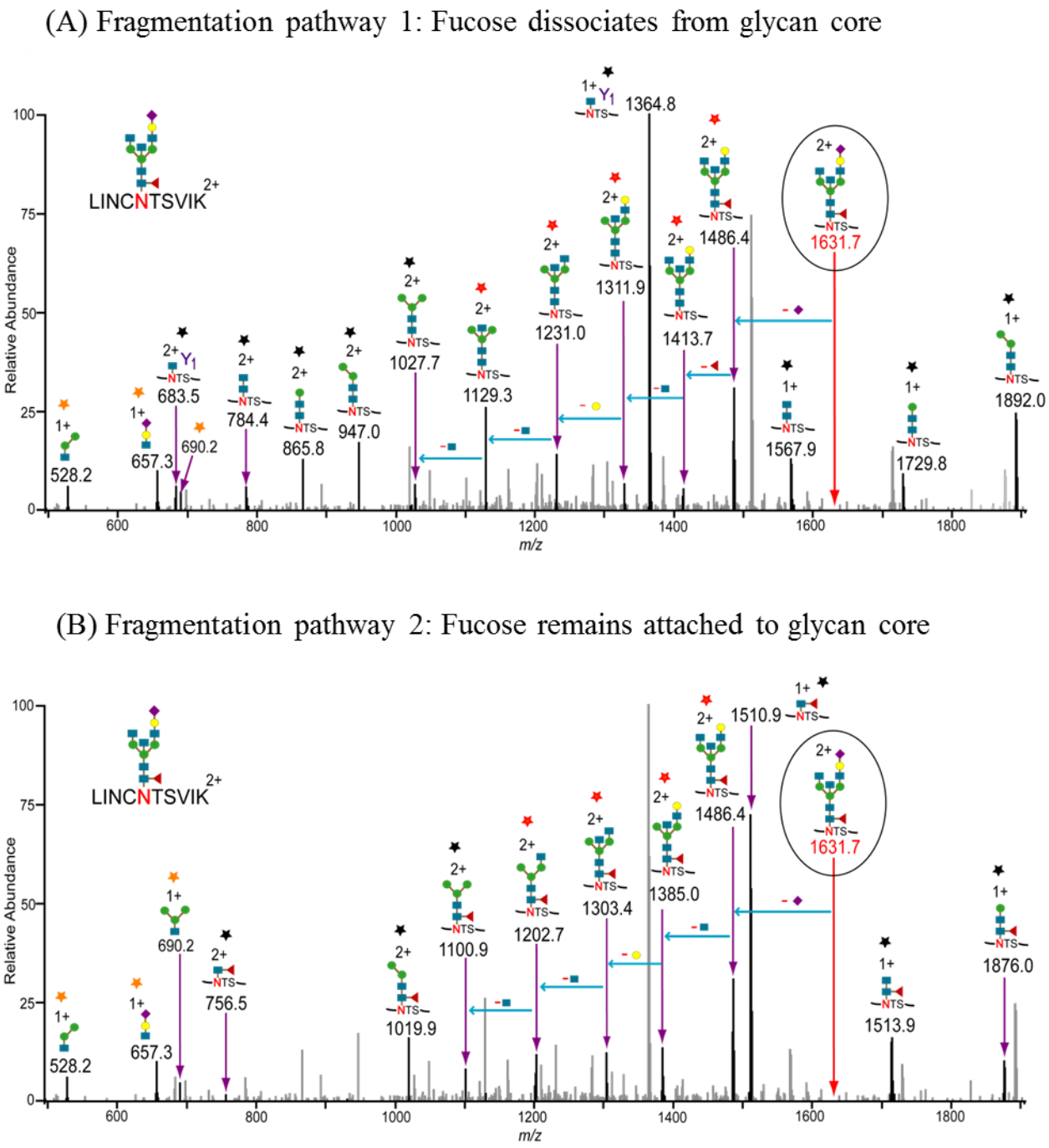
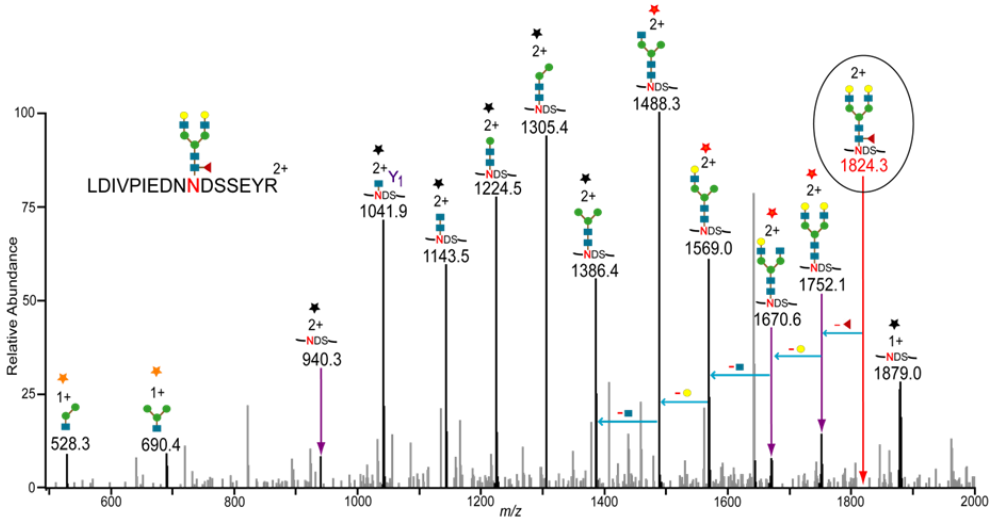


Figure 4. Representative CID spectra of a fucosylated complex type glycopeptide containing sialic acid showing two fragmentation pathways observed for fucosylated glycopeptides. **(A)** shows the fragmentation path that leads to [precursor – monosaccharide] product ions (red star) and [peptide + core component] product ions (black star) that do not contain the core fucose. **(B)** shows the fragmentation pathway that generates [peptide + core component] product ions (black star) and [precursor – monosaccharide] product ions (red star) that contain the appended core fucose. Oxonium ions (orange

star) are also observed. Ions generated by each fragmentation pathway are annotated and shown in black while non-related peaks are shown in gray. Precursor ions are in black ovals.

Analysis of CID data from fucosylated glycopeptides that do not contain sialic acid also revealed the same two fragmentation pathways described above. An example is shown in Figure 5.

(A) Fragmentation pathway 1: Fucose dissociates from glycan core



(B) Fragmentation pathway 2: Fucose remains attached to glycan core

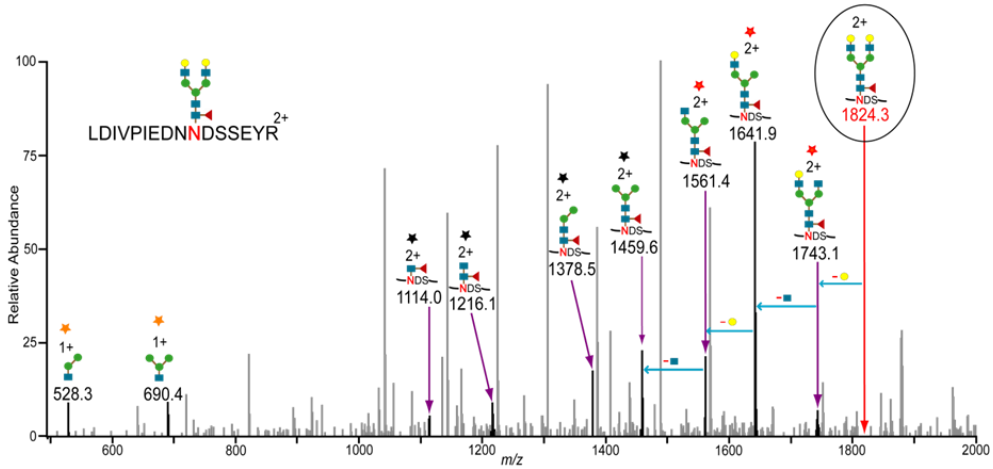


Figure 5. Representative CID spectra of a fucosylated complex type glycopeptide that does not contain sialic acid showing two fragmentation pathways observed for fucosylated glycopeptides. **(A)** shows the fragmentation path that leads to [precursor – monosaccharide] product ions (red star) and [peptide + core component] product ions (black star) that do not contain the core fucose. **(B)** shows the fragmentation pathway that generates [peptide + core component] product ions (black star) and [precursor – monosaccharide] product ions (red star) that contain the appended core fucose. Oxonium ions (orange star) are also observed. Ions generated by each fragmentation pathway are annotated and shown in black while non-related peaks are shown in gray. Precursor ions are in black ovals.

The characteristic [peptide + core component] and [precursor – monosaccharide] product ions from the two pathways were identified irrespective of the charge state of the precursor glycopeptide and the relative numbers of Hex and HexNAc residues of the glycan component.

After identifying the two common CID fragmentation pathways for fucosylated glycopeptides, we studied GlycoPep Grader's scoring rules to verify whether or not the software searches for the characteristic product ions (ions with and without appended fucose) generated by the two pathways when scoring CID data from fucosylated glycopeptide compositions. We noticed that the current version of GlycoPep Grader searches for product ions resulting from only one of the two possible fragmentation pathways: the product ions that do not contain appended fucose. Based on this finding, we propose that the algorithm be redesigned to search for additional product ions (product ions that contain the appended core fucose residue) to increase the scores of fucosylated glycopeptides compositions. The proposed ions are shown in Figure 6. The product ions (and one oxonium ion at *m/z* 366) that are searched for by the current version of GlycoPep Grader are shown in Figure 6A. Figure 6B includes the proposed product ions that contain core-fucose (ions indicated by red asterisks) that should also be searched for when scoring CID data of fucosylated glycopeptides. In addition, we also propose that the search

includes additional oxonium ions, as shown in the figure. The explanations for including the new oxonium ions are given in Section 4.4.1 of this chapter.

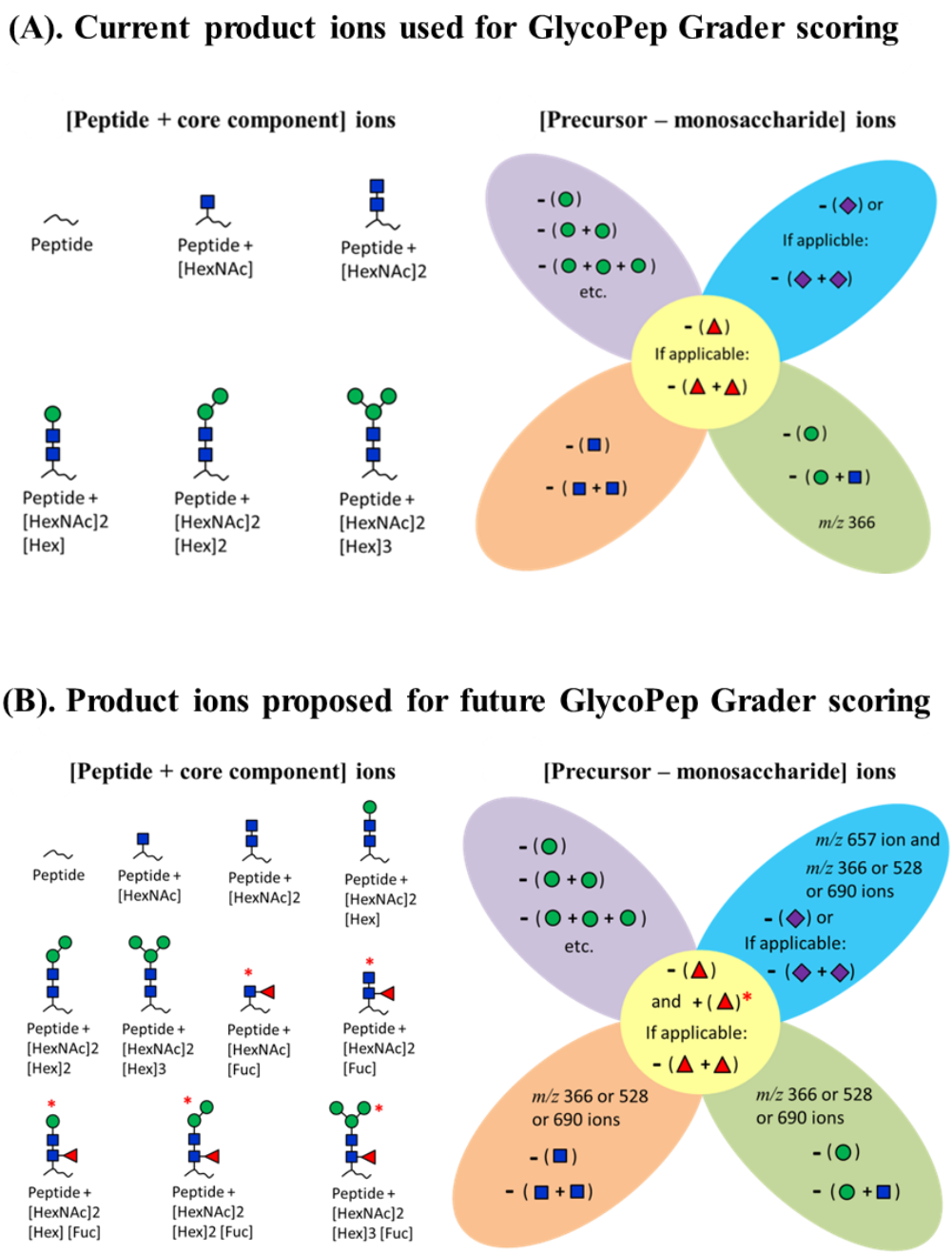


Figure 6. Schematic representation of the current product ions used by GlycoPep Grader to score *N*-linked glycopeptides **(A)** and proposed product ions for future GlycoPep Grader scoring **(B)**. The proposed product ions include ions containing appended fucose (ions with red asterisks), for the scoring of fucosylated glycopeptides. Prominent oxonium ions common in CID spectra of *N*-linked glycopeptides are also included. The color codes of the [precursor – monosaccharide] ions are explained in Figure 1. This figure is adapted from Reference 8.

4.4 GlycoPep Grader Updates

To determine whether the scores of fucosylated glycopeptides would increase by scoring product ions with appended fucose, we tested a rule that scores the [peptide + component] product ions containing fucose residues. For any fucosylated candidate glycopeptide, the software searched for [peptide + component] ions with or without appended fucose in them and gave credit for a correct assignment if either set of ions was identified. Representative results from the addition of this rule are shown in Table 1. The *PeptideScores* (highlighted in green) of some fucosylated glycopeptide candidates (targets1 and 2, and some decoys of target number 3) increased slightly, and this increase led to higher *GlycopeptideScores* (also highlighted in green). As shown in Table 1, the new rule helped to increase the scores of a few fucosylated glycopeptide candidates (both targets and decoys) but not most of them. This outcome was expected since our analysis of the peptide and glycan scores (shown in Figure 2) indicated that the glycan rules were more of a problem.

The new rules we proposed for scoring the [precursor – monosaccharide] product ions containing appended fucose, have not yet been implemented into the scoring algorithm, so this aspect of the work is ongoing. Yet, some scores in Table 1 indicate that the *GlycanScores* of fucosylated glycopeptide candidates can be increased drastically if the same rules for scoring non-fucosylated hybrid/complex glycopeptides are implemented for scoring fucosylated glycopeptides but with the addition of fucose residues to the product ions that are being searched for. For example, the glycan components of target 1 (glycan: [Hex]4[HexNAc]3[Fuc]1) and its decoy (glycan: [Hex]4[HexNAc]3) only differ by a fucose residue, but the non-fucosylated decoy has a *GlycanScore* of 1 (the highest possible glycan score for any glycopeptide) while the

target has a *GlycanScore* of zero. A similar phenomenon is observed for target number 3 and its first decoy. Hence, the *GlycanScores* of fucosylated glycopeptides can be drastically increased if the same rules for scoring non-fucosylated hybrid/complex glycopeptides are implemented but with the addition of fucose residues to the [precursor – monosaccharide] product ions that are scored.

Table 1. Example of glycopeptide candidate scores before and after addition of a new rule that searches for [Peptide + core component] ions with and without fucose.

Number	I.D	Peptide	PeptideScore		Glycan	GlycanScore		GlycopeptideScore	
			Old	New		Old	New	Old	New
1	Target	SNITGLLLVR	21 / 22 = 0.95	22 / 22 = 1	[Hex]4[HexNAc]3[Fuc]1	0 / 16 = 0	0 / 16 = 0	63.95	67.00
	Decoy	1212.66414	15 / 20 = 0.75	15 / 20 = 0.75	[Hex]4[HexNAc]3	8 / 8 = 1	8 / 8 = 1	83.25	83.25
2	Target	EEQFNSTFR	16 / 20 = 0.8	17 / 20 = 0.85	[Hex]5[HexNAc]4[NeuNAc]1[Fuc]1	0 / 12 = 0	0 / 12 = 0	53.6	56.95
	Decoy	1284.5635	13 / 18 = 0.72	13 / 18 = 0.72	[Hex]5[HexNAc]4[NeuNAc]1	4 / 8 = 0.5	4 / 8 = 0.5	64.89	64.89
3	Target	EEQYNSTYR	17 / 18 = 0.94	17 / 18 = 0.94	[Hex]3[HexNAc]5[Fuc]1	6 / 10 = 0.6	6 / 10 = 0.6	83.08	83.08
	Decoy	1316.56284	12 / 16 = 0.75	12 / 16 = 0.75	[Hex]3[HexNAc]5	6 / 6 = 1	6 / 6 = 1	83.25	83.25
	Decoy	1941.75504	2 / 6 = 0.33	3 / 6 = 0.5	[Hex]2[HexNAc]2[Fuc]1	0 / 8 = 0	0 / 8 = 0	22.33	33.50
	Decoy	643.26394	3 / 21 = 0.14	5 / 21 = 0.24	[Hex]5[HexNAc]6[Fuc]1	6 / 10 = 0.6	6 / 10 = 0.6	29.37	35.75
	Decoy	481.25254	3 / 20 = 0.15	5 / 20 = 0.25	[Hex]6[HexNAc]6[Fuc]1	0 / 12 = 0	0 / 12 = 0	10.05	16.75

Scores that improved due to the addition of the new rule are highlighted in green. *GlycanScores* that signal how the glycan scoring rules could be modified are highlighted in yellow.

4.4.1 Other Potential Areas for Improving GlycoPep Grader Scoring

After extensive study of GlycoPep Grader's scoring algorithm, we found additional areas where some changes can be made in order to improve the glycopeptide candidate scores, including those of fucosylated glycopeptides. First, the intensity threshold limit for scoring

product ions resulting from the cleavage of fucose and sialic acid residues is typically about three times higher than the intensity threshold for scoring the predominant product ions from non-fucosylated or sialylated glycopeptides. For example, at a normalization level of 2%, the intensity threshold limit for the predominant product ions from high-mannose and non-fucosylated glycopeptides is 3% while those from fucosylated glycopeptides is 10%. These levels were set on the basis that the fucose residue is very labile and product ions resulting from neutral loss of this residue would be very intense. However, we found that this is not always the case, as some spectra of fucosylated glycopeptides did not contain very intense [precursor – monosaccharide] product ions resulting from the loss of fucose from the precursor glycopeptide, and a few spectra completely lacked the ion. Hence, we propose that instead of using product ions resulting from the loss of fucose from the precursor glycopeptide ion as the predominant product ions for fucosylated glycopeptides, the Y1+Fuc ion should be used. This ion was present in all CID data of fucosylated glycopeptides and was the most intense ion containing fucose in some cases (e.g peak at *m/z* 1510.9 in Figure 4). In addition, we propose that the threshold limits for scoring product ions from fucosylated glycopeptide compositions should be the same as those from high-mannose and non-fucosylated hybrid/complex glycopeptide compositions.

Alternatively, a noise correction approach that prevents the use of different intensity threshold values for scoring different types of glycopeptides could be incorporated into GlycoPep Grader. CID data could be binned into 100 Da bins, and only the top 5 ions in each bin should be used for scoring glycopeptide candidates; we have shown previously that this approach is effective for scoring glycopeptides' ETD spectra.[16] In addition, the scoring algorithm could be changed so that the most intense peaks (e.g the top 10 most intense peaks) of each spectrum get

more weight, thus automatically penalizing any candidate for which all or some of the most intense peaks are not matched.

Another area where GlycoPep Grader can be improved is by increasing the scores of precursor ions at charge states of 3+ or higher. We found that although several precursor ions at a charge state greater than two out-score their decoys, the candidate scores are typically less than 45% while the scores of glycopeptides at charge state of two are mostly greater than 70%. Future updates to increase the score gap between decoy glycopeptides and target candidates at high charge states would greatly improve the scoring algorithm.

Finally, glycopeptide scores could also be increased if the scoring rules are modified so that the algorithm searches for more oxonium ions, as shown in Figure 6B. CID fragmentation of *N*-linked glycopeptides typically results in oxonium ions at *m/z* 366, 528, and 690[17,18] (see figures 4 and 5), and the spectra from glycopeptide compositions containing sialic acid have an additional oxonium ion at *m/z* 657[14,15] (see Figure 4). The current version of GlycoPep Grader only searches for the oxonium ion at *m/z* 366. This ion may be out of range for some data generated on ion trap instruments because the lowest mass range on these instruments is typically 1/3 the m/z value of the precursor ion.[19] Hence, instead of scoring only the oxonium ion at *m/z* 366, a new rule could be included to score the oxonium ion at *m/z* 657 (for sialylated glycopeptides only), and the ions at either *m/z* 366, or 528, or 690 (for all types of glycopeptide candidates). A higher weight could be given if at least two of these ions are matched.

4.5 Conclusion

The development of bioinformatics tools for the interpretation of glycopeptide MS/MS data is an emerging and challenging field. As more knowledge is being gained on the fragmentation behaviors of different types of glycopeptides of varying peptide lengths, and at different charge states, high quality software will be developed and more updates will be seen in the near future on the existing glycopeptide scoring algorithms to improve their glycopeptide scores.

In Chapter 3 of this dissertation, we showed that GlycoPep Grader has unprecedented strengths in scoring CID data from high-mannose and non-fucosylated hybrid/complex glycopeptide compositions, but the software does not effectively score spectra from fucosylated glycopeptide compositions. After extensive studies of GlycoPep Grader's scoring algorithm and the fragmentation characteristics of CID data from high-mannose, non-fucosylated hybrid/complex, and fucosylated hybrid/complex glycopeptide compositions, we identified an alternative fragmentation pathway for fucosylated glycopeptides that was not incorporated in GlycoPep Grader's scoring rules. The alternative fragmentation pathway leads to [Peptide + core component] and [precursor – monosaccharide] product ions that contain appended fucose residue. A test of a new GlycoPep Grader rule fore scoring [Peptide + core component] product ions containing fucose residues showed progress in improving the scores of fucosylated glycopeptides. The incorporation of the proposed rules for scoring [precursor – monosaccharide] product ions that contain appended fucose residue, and implementation of the additional proposed updates described in Section 4.4.1, should drastically improve the scores of fucosylated

glycopeptides and lead to no or few decoys out-scoring fucosylated glycopeptide target candidates.

4.6 References

1. Go, E. P.; Herschhorn, A.; Gu, C.; Castillo-Menendez, L.; Zhang, S.; Mao, Y.; Chen, H.; Ding, H.; Wakefield, J. K.; Hua, D.; Liao, H. X.; Kappes, J. C.; Sodroski, J.; Desaire, H. *J. Virol.,* **2015,** 89 , 8245-8257.

2. Desaire, H. *Molecular & Cellular Proteomics,* **2013,** 12 , 893-901.

3. Woodin, C. L.; Maxon, M.; Desaire, H. *Analyst,* **2013,** 138 , 2793-2803.

4. Hu, H.; Khatri, K.; Klein, J.; Leymarie, N.; Zaia, J. *Glycoconj. J.* **2016,** 33 , 285-296.

5. Hu, H.; Khatri, K.; Zaia, J., *Mass Spectrom. Rev.*, **2017**, 36, 475-498.

6. Lynn, K.-S.; Chen, C.-C.; Lih, T. M.; Cheng, C.-W.; Su, W.-C.; Chang, C.-H.; Cheng, C.-Y.; Hsu, W.-L.; Chen, Y.-J.; Sung, T.-Y. *Anal. Chem.,* **2015,** 87 , 2466-2473.

7. Chandler, K. B.; Pompach, P.; Goldman, R.; Edwards, N. *J. Proteome Res.*, **2013,** 12, 3652-3666.

8. Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. *Anal. Chem.,* **2012,** 84 , 4821-4829.

9. Hodoniczky, J.; Zheng, Y. Z.; James, D. C. *Biotechnol. Prog.,* **2005,** 21, 1644-1652.

10. Iida, S.; Misaka, H.; Inoue, M.; Shibata, M.; Nakano, R.; Yamane-Ohnuki, N.; Wakitani, M.; Yano, K.; Shitara, K.; Satoh, M. *Clin. Cancer Res.,* **2006,** 12, 2879-2887.

11. Jefferis, R. *Nat. Rev. Drug Discov.,* **2009,** 8, 226-234.

12. Vékey, K.; Ozohanics, O.; Tóth, E.; Jekő, A.; Révész, Á.; Krenyácz, J.; Drahos, L. *Int. J. Mass Spectrom.,* **2013,** 345–347, 71-79.

13. Tajiri, M.; Kadoya, M.; Wada, Y. *J. Proteome Res.,* **2009,** 8, 688-693.

14. Nwosu, C. C.; Seipert, R. R.; Strum, J. S.; Hua, S. S.; An, H. J.; Zivkovic, A. M.; German, B. J.; Lebrilla, C. B. *J. Proteome Res.,* **2011,** 10, 2612-2624.

15. Song, E.; Pyreddy, S.; Mechref, Y. *Rapid Comm. Mass Spectrom.,* **2012,** 26, 1941-1954.

16. Zhu, Z.; Hua, D.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.,* **2013,** 85, 5023-5032.

17. Conboy, J. J.; Henion, J. D. *J. Am. Soc. Mass Spectrom.,* **1992,** 3, 804-814.

18. Itoh, S.; Takakura, D.; Kawasaki, N.; Yamaguchi, T. *The Protein Protocols Handbook,* Walker, J. M., Ed. Humana Press: Totowa, NJ, **2009**; pp 1419-1430.

19. Lill, J. *Mass Spectrom. Rev.,* **2003,** 22, 182-94.

Chapter 5: Conclusion

5.1 Dissertation Summary

The work reported in this dissertation focused on disulfide bond characterization in IgG3 antibodies using tandem mass spectrometry and the development of a tool that assess the quality of algorithms that assign glycopeptides to tandem mass spectrometry data. Disulfide bonds and *N*-linked glycosylation are two important protein post-translational modifications (PTMs) that are common in proteins, and they play a vital role in maintaining protein stability, structure, and function.[1-4] As proteins are increasingly used for the development of biotherapeutics and as biomarkers for diseases, there is an accompanying need for thorough characterization of disulfide bonds and glycosylation, as well as other PTMs, in order to ensure the desired structural and biological properties of protein drugs and for proper identification of disease biomarkers. Tandem mass spectrometry is an invaluable tool for the characterization of these PTMs and for protein characterization in general.

High resolution mass spectrometry was used to study the disulfide connectivity of endogenous IgG3 antibodies to investigate whether or not they contain disulfide-mediated isoforms. Among the four classes of IgG's (IgG1, IgG2, IgG3, and IgG4), disulfide-based isomers have been identified for endogenous IgG2 and IgG4.[5-7] No disulfide isomers have been reported for endogenous IgG1, although low amounts of IgG1 disulfide isomers have been reported during drug development from recombinant IgG1.[8] Disulfide mediated isomers have not been studied for endogenous IgG3 due to its short half-life in serum, which made it not suitable

for drug development. However, IgG3 is now a viable drug candidate due to recent reports that show that the half-life of IgG3 can be extended by point mutation of amino acids in the CH3 domain.[9] This finding creates an urgent need for thorough characterization of the structural properties of IgG3 antibodies, such as its disulfide bond connectivity. By using electron transfer dissociation and collision induced dissociation, the native disulfide bond connectivity in the constant region of IgG3 was confirmed, but no alternative disulfide connectivity was identified, indicating that no disulfide isoforms are present in the constant region of endogenous IgG3.[10] Hence, alternative disulfide bonds should not be present in recombinant IgG3 molecules. This study provides the first benchmark for disulfide bond analysis in IgG3 using high resolution mass spectrometry.

Protein glycopeptide analysis is a challenging task due to the microheterogeneity of glycans. Several glycans can occupy the same different glycosylation site, making manual glycopeptide analysis from MS/MS data extremely challenging. As a result, numerous automated tools have emerged in recent years to facilitate glycopeptide analysis from MS/MS data.[11-13] This increase in the glycopeptide tools raises the need for a simple method to assess the quality of the tools in order to increase confidence in their results. However, there is no consensus in the proteomics field regarding the best method to assess an automated glycoproteomics tool. Two new tools, a *Glycopeptide Decoy Generator*, and a large dataset of manually assigned CID spectra from diverse glycopeptide compositions are reported in this dissertation. These tools enable end-users of glycoproteomics tools to rapidly assess the quality of any algorithm that assigns glycopeptides to CID data.[14] The tools can also be used by software developers to develop high quality scoring algorithms. The importance of the new tools was demonstrated by

using them to assess the quality of an existing software, GlycoPep Grader; a tool that assigns glycopeptide's CID data. We found that GlycoPep Grader has some weaknesses in scoring spectra from fucosylated glycopeptides. Finally, in Chapter 4, we investigated the source of the weaknesses in GlycoPep Grader and proposed new rules to improve its scoring algorithm.

Overall, the work reported in this dissertation presence a benchmark for future disulfide bond analysis in IgG3 antibodies,[10] and it introduces new tools that enable efficient glycopeptide software development and rapid evaluation of existing glycopeptide tools,[14] as well as a systematic approach to pin-point sources of flaws in glycopeptide scoring algorithms.

5.2 References

1. Zhang, L.; Chou, C. P.; Moo-Young, M., *Biotechnol. Adv.,* **2011,** 29 (6), 923-929.

2. Abkevich, V. I.; Shakhnovich, E. I., *J. Mol. Biol.,* **2000,** 300 (4), 975-985.

3. Xu, C.; Ng, D. T. W., *Nat. Rev. Mol. Cell Biol.,* **2015,** 16 (12), 742-752.

4. Liu, L., *J. Pharm. Sci.,* **2015,** 104 (6), 1866-1884.

5. Bloom, J. W.; Madanat, M. S.; Marriott, D.; Wong, T.; Chan, S. Y., *Protein Sci.,* **1997,** 6 (2), 407-415.

6. Wypych, J.; Li, M.; Guo, A.; Zhang, Z.; Martinez, T.; Allen, M. J.; Fodor, S.; Kelner, D. N.; Flynn, G. C.; Liu, Y. D.; Bondarenko, P. V.; Ricci, M. S.; Dillon, T. M.; Balland, A., *J. Biol. Chem.,* **2008,** 283 (23), 16194-16205.

7. Zhang, A.; Fang, J.; Chou, R. Y. T.; Bondarenko, P. V.; Zhang, Z., *Biochem.,* **2015,** 54 (10), 1956-1962.

8. Lu, C.; Liu, D.; Liu, H.; Motchnik, P., *mAbs,* **2013,** 5 (1), 102-113.

9. Stapleton, N. M.; Andersen, J. T.; Stemerding, A. M.; Bjarnarson, S. P.; Verheul, R. C.; Gerritsen, J.; Zhao, Y.; Kleijer, M.; Sandlie, I.; de Haas, M.; Jonsdottir, I.; van der Schoot, C. E.; Vidarsson, G., *Nat. Commun.*, **2011,** 2*:599.*

10. Lakbub, J. C.; Clark, D. F.; Shah, I. S.; Zhu, Z.; Su, X.; Go, E. P.; Tolbert, T. J.; Desaire, H., *Anal. Methods*, **2016,** 8 (31), 6046-6055.

11. Hu, H.; Khatri, K.; Klein, J.; Leymarie, N.; Zaia, J., *Glycoconj. J.*, **2016,** 33 (3), 285-296.

12. Hu, H.; Khatri, K.; Zaia, J., *Mass Spectrom. Rev.*, **2017**, 36 (4), 475-498.

13. Woodin, C. L.; Maxon, M.; Desaire, H., *Analyst*, **2013,** 138 (10), 2793-2803.

14. Lakbub, J. C.; Su, X.; Zhu, Z.; Patabandige, M. W.; Hua, D.; Go, E. P.; Desaire, H., *J. Proteome Res.*, **2017**, doi:10.1021/acs.jproteome.7b00289.