Software for Automated Interpretation of Mass Spectrometry Data from Glycans and Glycopeptides

Carrie L. Woodin, Morgan Maxon, and Heather Desaire*
Department of Chemistry, University of Kansas, Lawrence, KS

Abstract

The purpose of this review is to provide those interested in glycosylation analysis with the most updated information on the availability of automated tools for MS characterization of N-linked and O-linked glycosylation types. Specifically, this review describes software tools that facilitate elucidation of glycosylation from MS data on the basis of mass alone, as well as software designed to speed the interpretation of glycan and glycopeptide fragmentation from MS/MS data. This review focuses equally on software designed to interpret the composition of released glycans and on tools to characterize N-linked and O-linked glycopeptides. Several websites have been compiled and described that will be helpful to the reader who is interested in further exploring the described tools.

INTRODUCTION

Overview of Glycosylation

The addition of monosaccharide residues onto a protein or lipid, known as glycosylation, serves an important function in many cellular signaling and communication events, including those involving host-pathogen interactions.1–4 It has long been understood that protein-carbohydrate interactions play a participatory role in many processes affecting disease progression.1, 4–7 Furthermore, experimental evidence demonstrates that the identity of the attached glycans change during these events.6, 8–10 For example, aberrant glycosylation is often present in individuals experiencing cancer, diabetes, and inflammation.4, 8–11 As such, accurate characterization of a glycoprotein’s glycan substituents has been shown to be crucial in the development of potential biomarkers, protein-based vaccine candidates, and pharmaceutical treatments.5, 9, 12, 13

In studies involving protein glycosylation, mass spectrometry (MS) is currently considered the most utilized instrument for the characterization of glycan substituents, both attached or released.11 Accordingly, this review focuses on those automated analysis tools that can be applied to more efficiently analyze glycosylation profiles through interrogation by MS and MSⁿ methods.
Glycosylation Heterogeneity

Unlike DNA replication and protein transcription, glycosylation is a “non-template”-driven process,5,14 where the sugar residues form a multitude of arrangements.5,12 The monosaccharides that comprise the glycan may be long or short, branched or linear, and linked in a variety of ways, creating a large degree of variability.3,5,12 This heterogeneity is described in two ways: Glycan differences at different sites of attachment (macroheterogeneity), or within the same site (microheterogeneity).12 The large amount of heterogeneity presents a challenging obstacle to researchers attempting to elucidate structural, as well as compositional, information on a protein’s glycan population, especially when samples are mixtures of proteins.

Types of Protein Glycosylation

Over half of all proteins expressed are predicted to be glycosylated.15 In addition to the established forms of protein glycosylation, including N-linked, O-linked, and C-linked forms,1,16–18 rarer configurations such as S-linked glycosylation, are also being discovered.19 Although a variety of types exist, the two most common types of glycosylation are N-linked and O-linked.1,7,20 In N-linked glycosylation, the addition of a glycan may occur at the asparagine residue when the consensus sequence Asn-Xaa-Ser/Thr occurs, where Xaa is any amino acid except proline.12,18,21 The inclusion of this pattern is a fundamental requirement for N-linked glycosylation to occur, though it is not a guarantee that a glycosylation site will be occupied.21 With O-linked glycosylation, the glycan addition may occur at any Ser or Thr residue within the protein sequence,4,11 though a very low percentage of these sites are actually occupied.15 In contrast to N-linked glycans, O-linked glycans have less defined sequence patterns,12,18,22 and may consist of several distinct core arrangements.11 For these reasons, both the prediction and determination of O-linked glycosylation characteristics has advanced slower than N-linked glycosylation analysis.11

Glycosylation Analysis

In the determination of glycan structural information, mass spectrometry (MS) has shown to be a powerful tool, as a large part of the successful interrogation of glycan composition and structures has been accomplished utilizing MS experiments.18,23,24 There are two main strategies for elucidating glycosylation information using MS techniques: 1) Characterization of a protein’s glycans after they are released from glycoproteins and 2) Characterization of glycopeptides after proteolytic digestion of a glycoprotein.11,18,23 The study of released glycans is particularly useful when rapid analysis of glycan composition is desired. Though N- and O-linked glycan populations can be studied independently through the use of different cleavage procedures,11,12 no information on where the individual glycans were attached along the protein is obtained when the glycans are cleaved a priori. In order to obtain glycosylation site-specific information for individual glycoforms, the second method, glycopeptide analysis, which requires digestion of the protein using a protease such as trypsin, is necessary.18,23,25 This method is generally advantageous because it provides information about both glycan composition and the site of the glycan’s attachment.23 Despite the challenges mentioned in this review, techniques that allow for complete
profiling of a peptide’s glycan population have advanced greatly in the past decade, especially with respect to site-specific glycopeptide analysis.

In the study of protein glycosylation by either of these techniques, a number of resources, including databases providing information on known glycosylation structures or site occupancy, as well as collections of experimental data, are currently available. For instance, researchers needing to identify occupied N-linked glycosylation sites on a specific protein can access UniProtKB, while those wanting statistics specific to proteins modified by O-GlcNAc could visit dbOGAP. Although this repertoire of information is greater for proteins modified by N- and O-linked glycan types, databases that contain entries on Cglycosylated proteins, such as dbPTM, are available as well. A current list and description of these database resources are provided in Table 1.

**AUTOMATED ANALYSIS OF RELEASED GLYCANS**

**Characterization of High Resolution MS data for Oligosaccharides**

Often, the easiest way to identify a protein’s glycan population is by enzymatically cleaving the glycan substituents and analyzing the monosaccharide residues directly. N- and O-linked glycans from the same protein can be independently characterized in this manner, as in the method described by Goetz et al. where β-elimination is used to release O-linked glycans, which are simultaneously permethylated. Once cleaved, automated analysis tools that determine glycan composition from MS data may be used.

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

Another software solution useful for the identification of glycans from MS data is SysBioWare, described by Vakhrushev et al. SysBioWare takes raw MS data that a user uploads and performs baseline adjustment and denoising, wavelet analysis, and peak detection before grouping isotopes of detected peaks. The isotopic grouping is also performed automatically, which enables the program to deduce monoisotopic m/z values and precursor charge states without the need of manual input by the user. The software then determines monosaccharide compositions on the basis of mass. Currently, the SysBioWare program is being updated to include analysis of MS/MS data for glycans as well. SysBioWare is not freely available to the public at this time.
Similar to SysBioWare is GlycoWorkbench. GlycoWorkbench evaluates glycan compositions (which are proposed by the user) by searching the spectral peak list of user-input MS data for matches between calculated theoretical glycan masses and corresponding m/z values.\(^3\) The GlycanBuilder tool, designed to interface with GlycoWorkbench, enables the drawing of glycan structure representations, with all stereochemical information on the monosaccharides depicted as specified by the user.\(^4\) Both analysis tools, GlycanBuilder and GlycoWorkbench, are available online free of charge, as described in Table 2.

GlycoSpectrumScan, another freely available program, was developed by Deshpande et al.\(^1\) and works to identify N- and O-linked glycoforms using MS\(^{1}\) data.\(^4\) This software is capable of analyzing both singly or multiply charged ions directly from raw data, and accepts the input of both ESI and MALDI spectra.\(^4\) GlycoSpectrumScan also determines the relative abundance of N- and O-linked glycoforms that are identified for each glycosylation site.\(^4\) However, the user must enter the N- and/or O-linked glycan compositions potentially present in the sample, as well as the \textit{in silico} peptide masses of the digested glycoprotein.\(^4\) GlycoSpectrumScan is available online (see Table 2).

**MS/MS Approaches for Glycan Characterization**

Until recently, when automated software tools and scoring algorithms became available, the identification of accurate glycan or glycopeptide assignments from MS/MS data was a key bottleneck, due to the need for extensive manual data analysis. STAT, designed by Gaucher et al.,\(^4\) is one of the first automated tools for the determination of glycan composition using tandem MS.\(^4\) STAT is designed for glycans of up to ten monosaccharide residues, and has the ability to quickly analyze relevant N-glycan compositions.\(^4\) STAT also lists the most likely structures in order of probability to provide a ranking system when more than one candidate glycan matches the fragmentation profile of the data being analyzed.\(^4\) Unfortunately, this program is no longer publicly accessible.

An early analysis tool capable of evaluating O-linked glycan fragmentation is the OSCAR algorithm.\(^4\) OSCAR, as developed by Ashline et al.,\(^4\) is specifically designed for the annotation of permethylated O-linked oligosaccharides from MS\(^{n}\) data.\(^4\) OSCAR is part of a collection of software tools termed Glyspy, which is not currently accessible to the public.\(^4\) Although innovative, the use of OSCAR is limited to direct infusion experiments, as the software cannot effectively process data from LC-MS methods.\(^4\)

A program contemporary to OSCAR and also developed to handle glycan MS/MS data is StrOligo.\(^4\) This instrument-specific program was designed by Ethier et al. for the determination of N-linked glycan structures from MALDI MS/MS data.\(^4\) In published research, StrOligo successfully assigned the correct glycan structure in 24 out of 28 cases.\(^4\) Although the results of these two programs are promising, neither program is freely accessible online.

Several alternative glycan analysis tools are freely available online. One of the earliest of these was reported by Lohmann et al. in 2004.\(^5\) The authors describe the web tools GlycoFragment and GlycoSearchMS, which were developed for glycan structural determination.\(^5\) The theoretical fragmentation patterns of carbohydrate structures are
calculated using GlycoFragment, which displays theoretical b- and y-fragments as well as c-, z-, a- and x-ions.\textsuperscript{45} GlycoSearchMS works to analyzes experimental glycan data by comparing it against a library of theoretical spectra from N-linked and O-linked glycan fragmentation entries extracted from SweetDB.\textsuperscript{45} The GlycoFragment program has been validated on both N-linked and O-linked glycan classes, and, used in conjunction with GlycoSearchMS, enables researchers to determine the most probable glycan composition according to the information from the combined algorithms.\textsuperscript{45,46} Both GlycoFragment and GlycoSearchMS are freely available. See Table 2 for more information.

Another heavily used, free, online tool for glycoform analysis is GlycoWorkbench, which has shown to be a resourceful tool not only for analysis of MS\textsuperscript{1} data, as mentioned previously, but in the identification of glycans from MS/MS data as well.\textsuperscript{39} To utilize the glycan fragmentation analysis feature, a user must first input/define the possible glycan compositions and spectral peak list.\textsuperscript{39} The software then calculates expected glycan fragmentation and relative \textit{m/z} values, and annotates peaks of the uploaded data with the most probable identity (shown in red to distinguish it), of all compositions tested.\textsuperscript{39} As previously stated, GlycoWorkbench is available for free online.

In addition to the freely available tools mentioned above, several other MS/MS analysis tools for glycans are available for researchers, either for purchase or by special request to the tools’ developers. Two of these are GlyCH and Glyquest.\textsuperscript{47,48} GlyCH was developed by Tang \textit{et al.} to perform automated interpretation of oligosaccharide tandem mass spectra.\textsuperscript{47} The algorithm has a scoring function built in to allow researchers to compare compositions when more than one is determined to be possible.\textsuperscript{47} The GlyCH algorithm, which has so far been tested on released N-glycans, is also capable of \textit{de novo} analysis, providing no more than 10 monosaccharide residues comprise the glycan chain.\textsuperscript{47} Although not freely accessible online; the program is available upon request from the authors.\textsuperscript{47} More recently, Gao \textit{et al.} developed Glyquest, an automated analysis program that takes a different approach to determine compositions of intact N-linked glycans.\textsuperscript{48} This software utilizes a database in conjunction with an integrated search engine to determine the composition of peptide-attached N-glycans from CID MS/MS data.\textsuperscript{48} After the program algorithmically determines the molecular weight of the protonated peptide within a given spectrum; candidate N-glycan compositions are selected and fragmented \textit{in silico} to generate a theoretical spectrum that is then compared to the experimental spectrum.\textsuperscript{48} The glycan compositions with fragmentation profiles that are most similar to the experimental fragmentation are determined to be the most probable candidates.\textsuperscript{48} Glyquest is not freely available to the public.

SimGlycan is another tool that can be used to increase throughput of glycan analysis.\textsuperscript{49,50} More information is available online (http://www.premierbiosoft.com/). This for purchase tool is useful for determining glycan structures from MS/MS data obtained from many different mass spectrometers, once an acquisition file is converted into mzXML format.\textsuperscript{50} A user uploads an MS/MS data file, and the software utilizes a built-in database with theoretical fragmentation profiles of nearly 10,000 glycan structures to provide the most likely structural candidates.\textsuperscript{49} One unique feature of SimGlycan is that no filtering of biologically relevant structures is provided, which can be advantageous for identifying novel
glycan structures, but disadvantageous in that it returns a user many structures which are not relevant.\textsuperscript{49} However, commercially available software such as SimGlycan are expensive, which potentially limit their use.

A more recent program developed specifically for the compositional interpretation of \(O\)-linked glycan fragmentation is CartoonistTwo, as described by Goldberg \textit{et al.}\textsuperscript{36} CartoonistTwo was designed using CID data acquired on an FTICR-MS, and validated using data from a test set of 34 spectra acquired from \textit{Xenopus} egg jelly.\textsuperscript{36} Unfortunately, the program is not freely accessible to the public.

**AUTOMATED ANALYSIS OF GLYCOPEPTIDES**

For researchers performing site-specific glycosylation analysis, the initial step toward accomplishing the characterization of attached glycoforms at unique sites within a digested protein is to identify potential glycosylation sites within that protein; the tools to facilitate this step are described in Table 1 of this review. In addition to these, programs that utilize algorithms to predict the likelihood of site-occupancy by examination of the amino acid residues surrounding the potential glycosylation site have also been developed.\textsuperscript{51–59} These prediction tools, along with a description and link to each tool, are provided in Table 3.

**Experimental Data Requirements**

After the resultant glycopeptides are obtained from the proteolytic digest, two types of data are generally used to accurately characterize the identity of a glycopeptide. First, high resolution MS data of the glycopeptide is used to infer possible glycopeptide compositions; second, tandem MS data is acquired to distinguish between isomers and isobars.\textsuperscript{60} Figure 1 provides a schematic of this analysis process.

**\(N\)-LINKED GLYCOPEPTIDES**

Although \(N\)-linked glycoforms share a common core structure, the rest of the glycan follows one of three distinct arrangements. Based on the arrangement pattern, \(N\)-linked glycans compositions are classified into three main types, those with: 1) High mannose type glycans 2) Complex type glycans and 3) Hybrid type glycans.\textsuperscript{12,18} This information is useful when deciphering glycopeptide compositions from MS experiments, specifically from CID MS/MS data.\textsuperscript{61}

**\(N\)-linked Glycopeptide MS Data**

A variety of automated and semi-automated analysis tools have been created to aid in the interpretation of \(N\)-linked glycopeptide MS data. The key objective of these tools is to provide glycopeptide compositions that are consistent with the high resolution MS data. Researchers then typically use MS/MS analysis, described later in this review, to determine which of the compositions is correct for each given ion. Three of these tools are accessible to the public: GlycoMod (http://web.expasy.org/glycomod/), GlycoPep DB (http://hexose.chem.ku.edu/glycop.htm), and the previously mentioned GlycoSpectrumScan.\textsuperscript{41,62,63} GlycoMod, the earliest and most heavily used tool, accepts a protein sequence, possible monosaccharide building blocks, and experimental mass data as inputs, and it calculates all
possible glycopeptide compositions that fall within the mass tolerance. One restriction in the capacity of GlycoMod to analyze glycopeptide data is the inability to handle multiply-charged precursors.

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

**N-linked Glycopeptide MS/MS Data**

Each common N-linked glycan type (complex, hybrid, or high mannose) has a signature fragmentation profile that is present when a glycopeptide is subjected to MS/MS experiments. These characteristic fragmentation profiles are useful for determining the correct composition of an N-linked glycopeptide when isobaric candidate compositions are possible. However, as manual interpretation of these data are challenging, software is required to speed analysis time.

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

Many other publicly available tools to elucidate glycosylation profiles of glycopeptides have also emerged out of glycan analysis software. GlycoWorkbench and Glyco-Peakfinder both can annotate glycan fragmentation in glycopeptide data, although the peptide portion of the glycopeptide must be determined by some means other than the use of these tools. On the positive side, Glyco-Peakfinder is useful for de novo calculation and annotation of glycan fragment ions within tandem mass spectra. Users may allow constraints on the oligosaccharide such as size and attachment of other substituents (such as acetate, phosphate, and sulfate), and the program is capable of annotating multiply-charged ions (−4 to +4). Additionally, glycan fragmentation is analyzed across multiple charge states, and across multiple charge carriers (cationic carriers), within the same spectrum.
A completely different approach is used in GlycoPep ID.\textsuperscript{65} GlycoPep ID is a web-based tool developed by Go \textit{et al.} to interpret MS/MS data of glycopeptides and to identify the peptide component of glycopeptides through analysis of expected product ions.\textsuperscript{65} The URL to access this program is listed in Table 4. Although this program is useful for identification of the peptide portion of the glycopeptide in complex LC-MS samples, it does not contain a scoring algorithm to identify the most probable glycopeptide match.\textsuperscript{65}

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

Some alternative, unique strategies have been developed with the goal of scoring MS/MS data against potential glycopeptide compositions, such as those described using Peptoonist, Medicel Integrator, the Branch-and-Bound algorithm, GlycoMaster, Sweet Substitute, and GlyDB.\textsuperscript{66,67,68,69,70,71} Unfortunately, none of these programs are currently publicly available.

To address the need for publicly accessible tools specifically designed to interpret and score fragmentation of glycopeptides, GlycoMiner was developed by Ozohanics \textit{et al.}\textsuperscript{72} In the analysis of 3132 spectra, the program was reported to have found 338 that corresponded to MS/MS data of glycopeptides (versus peptides).\textsuperscript{72} Designed using qTOF data, the software is capable of assigning glycopeptide compositions when both the peptide and glycan components portions are unknown.\textsuperscript{72} However, the program is only capable of performing compositional analysis when the spectra are of good quality.\textsuperscript{72} The program fails when spectral quality is low, as evidenced by the program’s identification of glycan composition in only 196/338 glycopeptide spectra.\textsuperscript{72} Although this tool is a great advancement towards automated interpretation of glycopeptide MS/MS data, GlycoMiner often generates multiple plausible compositions and fails to rank the correct glycopeptide as the top candidate.\textsuperscript{72} In addition, the program requires spectra containing a low S/N ratio as well as the presence of low-mass marker ions, which are not typically present in data collected on ion trap instruments.\textsuperscript{72} Available online, GlycoMiner is free to download and use; see Table 4.

Similar to GlycoMiner, GlycoPeptide Search (GPS) is a recently developed program by Chandler \textit{et al.} for the determination of glycopeptide composition from CID data.\textsuperscript{73} Designed for purified glycoprotein samples analyzed by LC-MS/MS, GPS utilizes GlycomeDB, a glycan database in conjunction with the peptide file, which is supplied by the user, to generate an Excel file of glycopeptide matches based on fragmentation evidence.\textsuperscript{73} To generate the peptide-glycan pairs, GPS must find both low mass oxonium, and N-glycan core-containing, product ions.\textsuperscript{73} GPS is freely available online as well.\textsuperscript{73} For further information, see Table 4.
The targeted MS/MS approach utilizing the computational tool GlypID recently described by Wu et al. aims to characterize N-linked glycopeptides through the combined use of MS\(^1\) and MS\(^2\) information extracted from LC-MS/MS experiments.\(^{74}\) One of the benefits to the method is that no prior knowledge of the potential glycosylation or identity of the glycopeptide is necessary.\(^{74}\) Instead, GlypID assigns a cluster of glycopeptides in the “same family” (microheterogeneities) based on observed mass.\(^{74}\) In addition, the approach utilizes an isotope deconvolution algorithm to assign ion charges along with monoisotopic ions.\(^{74}\) This information is then added to the inclusion list of “prioritized precursor ions” for the MS/MS analysis that follows.\(^{74}\) Next, the resultant CID data is searched for the longest series of glycosidic bond cleavage series.\(^{74}\) These product ions are used to determine the *oligosaccharide sequence tag*, which is used to verify whether or not the spectrum is from a glycopeptide.\(^{74}\) A score is assigned to the CID spectrum based on this sequence tag.\(^{74}\) MS data is used to evaluate and score the relative probability of a glycopeptide by examining the clusters of peptide glycoforms, or those glycopeptides with the same peptide backbone that co-elute within a specific time range.\(^{74}\) The glycoform is then identified using the mass of the attached N-linked glycan, though the most current version of GlypID allows the entry of user-defined glycan compositions as well.\(^{74}\) A limitation to the program is that when low resolution data is used, there is a significant increase in the number of false-positive identifications of glycopeptide microheterogeneities within a cluster. Although the new targeted MS/MS approach has been optimized for FT-MS instrumentation and data, the original GlypID algorithm was designed using LC-MS ion trap data.\(^{75}\) A publicly accessible version of the computational tool is currently available online, free of charge to users (see Table 4).

Mayampurath et al. have now modified the GlypID algorithm with a scoring function that works to determine glycopeptide composition from high-energy C-trap dissociation (HCD) MS/MS data.\(^{76}\) The new software tool, GlypID 2.0, uses high resolution MS\(^1\) data along with CID and HCD scan information to improve the accuracy of N-linked glycopeptide identification.\(^{76}\) Like the original GlypID, GlypID 2.0 also works to score CID spectra independently on MS systems that do not contain the HCD instrument option.\(^{76}\) GlypID 2.0 is freely available to download, as listed in Table 4.

Woodin et al., have also developed a freely accessible web-based tool, GlycoPep Grader (GPG), to assign a glycopeptide composition to MS/MS data in an automated fashion.\(^{61}\) This tool is specifically designed for data collected in an ion trap mass spectrometer, and it features a novel algorithm that enables users to identify the correct glycopeptide composition from a pool of candidate compositions of the same nominal mass.\(^{61}\)

GPG utilizes the MS/MS data by calculating, scoring, and searching for the expected product ions of potential glycopeptide candidate compositions.\(^{61}\) The algorithm scores the glycopeptide candidate composition through detection of two types of product ions: 1) Ions that contain the peptide portion and some portion of the pentasaccharide core, or [peptide + core component] ions, and 2) Ions formed via neutral loss of monosaccharide residues from the precursor ion, or [precursor – monosaccharide] ions.\(^{61}\) The algorithm that powers GPG has been shown to assign the correct glycopeptide candidate after performing the MS/MS peak list search with a very high degree of accuracy.\(^{61}\)
One advantage to the algorithm behind GPG is that the precursor ion’s charge state is included in the input data, so all product ions can be searched for in a charge state specific manner.\textsuperscript{61} Secondly, no spectral transformation (to singly charged ions) needs to be performed prior to using the program as GPG automatically searches for product ions in a charge-specific fashion, bypassing the need for additional processing software.\textsuperscript{61} A disadvantage of the program is that the user must utilize a separate program, such as GlycoMod, to obtain potential matches for the high resolution MS data, prior to assigning the MS\textsuperscript{2} data with GPG.\textsuperscript{61} GPG can be found online, and is free to use.

\section*{O-LINKED GLYCOPEPTIDES}

The analysis of O-linked glycoforms is particularly challenging, as no single consensus sequence exists to predict the site of glycan attachment.\textsuperscript{7,77} Further adding to the difficulty of analysis, factors that affect the efficiency of glycosylation at N-linked sites are different than those affecting O-glycosylation efficiency. For example, the presence of aromatic residues near an O-linked site inhibits glycosylation; whereas the presence of an aromatic residue near an N-linked site increases the likelihood of site-occupancy.\textsuperscript{22}

\subsection*{Mucin-type O-linked Glycosylation}

The most prevalent form of O-linked glycosylation to occur in eukaryotic organisms is mucin-type O-glycosylation, which occurs where glycans are attached to a protein by the addition of \(\alpha-N\)-acetylgalactosamine (GalNAc) residues to the hydroxyl group of Ser/Thr side chains (commonly referred to as the Tn antigen).\textsuperscript{2,7} Though still in the infancy stage, analysis tools have recently been created to assist researchers in the determination of O-linked glycoforms, many of which are mucin in type, from MS data.

\subsection*{O-linked Glycopeptide Characterization from MS Data}

Recently, Deshpande \textit{et al.} advanced the MS data analysis of N- and O-linked glycopeptides with the advent of the GlycoSpectrumScan program.\textsuperscript{41} GlycoSpectrumScan is designed to analyze LC-MS data of intact glycopeptides from proteolytic digests.\textsuperscript{41} The program utilizes MS\textsuperscript{1} data to determine glycopeptide composition, along with the relative distribution of glycoforms at each of the sites.\textsuperscript{41} In addition, the algorithm behind the program offers a few distinct advantages in that it handles multiply charged ions, making it amenable to both MALDI and ESI data, and is currently freely available online (\texttt{www.glycospectrumscan.org}).\textsuperscript{41}

GlycoX and GlycoMod, described earlier in the analysis of N-linked glycopeptides, are capable of O-linked glycopeptide data interpretation as well.\textsuperscript{52,78} Unlike GlycoMod (\texttt{http://web.expasy.org/glycomod/}), GlycoX is not publically-available, though it is available upon request from the authors.\textsuperscript{78} GlycoWorkbench, also described previously, performs automation of O-linked glycopeptide MS\textsuperscript{1} data to elucidate the most likely composition from an experimental peak list in the same manner for N-linked glycopeptide spectra.\textsuperscript{39} GlycoWorkbench is freely available online (\texttt{http://download.glycoworkbench.org}).\textsuperscript{39}
O-linked Glycopeptide Characterization from MS/MS Data

Currently, there are no freely available stand-alone programs designed to automate the analysis of O-linked glycopeptide CID MS/MS data through evaluation of both unknown portions of a glycopeptide, the peptide and glycan. The GlycoWorkbench program is capable of annotating glycans in CID fragmentation data of glycopeptides. However, as described for the MS/MS characterization of N-linked glycopeptides, the identity of the peptide portion must already be known, as GlycoWorkbench solely evaluates the fragmentation of the glycan-containing portion of a glycopeptide.39

There are promising advances being made in the compositional determination of glycopeptides using electron transfer dissociation (ETD) fragmentation techniques, or a combination of CID and ETD, particularly in the study of O-linked species. A recent method described by Daracula et al. in which MS1, CID, and ETD data are used in conjunction with Protein Prospector v5.3 for the identification of SA1-10GalGalNAc-containing O-linked glycopeptides enriched from bovine serum, demonstrates the potential for automated analysis through a combination of these techniques and database searches. However, this process is only semi-automated and works only in the case of simple carbohydrate structures are present in the sample. Hopefully, the compositional information gained between the two complementary fragmentation methods of CID and ETD will enable researchers to develop algorithms and gain insight into creating automated programs to speed the analysis of O-linked MS/MS glycopeptide data as well.

CONCLUSION

For the study of protein glycosylation, there are two main approaches used by researchers: Glycan analysis and glycopeptide analysis. The less challenging mode of analysis is to release the glycans from a glycoprotein and analyze them independently. However, the most informative approach is to utilize a protease and cleave the glycoprotein into glycopeptides, retaining information on where each glycan is attached within the protein sequence. The generation of automated MS and MS/MS analysis tools to assist in the characterization of both glycans and glycopeptides is emerging as an effort to facilitate more rapid characterization of glycosylation by either method. Without these analysis tools, data interpretation is a difficult and tedious task. Currently available and recently developed programs for various glycan types, along with their associated advantages and disadvantages, are discussed within this review. With access to the automated tools described here, research in glycomics and glycoproteomics can be greatly facilitated.

Acknowledgments

The authors acknowledge financial support from the National Institutes of Health (RO1RR026061) and an NSF Career award (0645120) to HD, an NSF Fellowship (DGE-0742523) and Pfizer Award to CW, and a Seo Scholarship to MM.

References


Figure 1.
Flow chart outlining the use of MS and MS/MS data for glycopeptide identification.
### Table 1

Glycosylation Databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>Link to Database</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProtKB</td>
<td><a href="http://www.uniprot.org/">http://www.uniprot.org/</a></td>
<td>N-glycos. O-glycos. C-glycos.</td>
<td>Contains annotation of N-, O-, and C-linked glycosylation, as well as glycation. Both mammalian and non-mammalian entries are provided.</td>
</tr>
<tr>
<td>dbPTM</td>
<td><a href="http://dbptm.mbc.nctu.edu.tw/">http://dbptm.mbc.nctu.edu.tw/</a></td>
<td>N-glycos. O-glycos. C-glycos.</td>
<td>Contains a combinational repertoire of protein PTMs from other databases, including experimentally obtained data on site of modification.</td>
</tr>
<tr>
<td>O-GlycBase</td>
<td><a href="http://www.cbs.dtu.dk/databases/OGLYCBASE/">http://www.cbs.dtu.dk/databases/OGLYCBASE/</a></td>
<td>O-glycos. C-glycos.</td>
<td>Contains over 2000 entries of protein glycosylation sites, the majority of which are O-linked.</td>
</tr>
<tr>
<td>GlycoBase</td>
<td><a href="http://glycobase.univ-lille1.fr/base/">http://glycobase.univ-lille1.fr/base/</a></td>
<td>N-glycos.</td>
<td>Contains HPLC elution positions for 2-AB labeled N-linked glycans from exoglycosidase sequencing and LC-MS data.</td>
</tr>
</tbody>
</table>
Table 2
Tools available online to facilitate glycan characterization from MS and MS/MS data.

<table>
<thead>
<tr>
<th>MS Analysis Tool</th>
<th>Link to Tool</th>
<th>Concept and Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlycoWorkbench</td>
<td><a href="http://download.glycoworkbench.org/">http://download.glycoworkbench.org/</a></td>
<td>Identifies and annotates MS and MS/MS data with appropriate glycan compositions or fragments.</td>
</tr>
<tr>
<td>GlycoFragment</td>
<td><a href="http://www.glycosciences.de/tools/GlycoFragments/fragment.php4">http://www.glycosciences.de/tools/GlycoFragments/fragment.php4</a></td>
<td>Identifies and displays the main product ions expected for oligosaccharide MS/MS data.</td>
</tr>
<tr>
<td>GlycoSearchMS</td>
<td><a href="http://www.glycosciences.de/database/start.php?action=form_ms_search">http://www.glycosciences.de/database/start.php?action=form_ms_search</a></td>
<td>Compares experimental MS/MS data to product ions calculated from an extensive library of N- and O-linked glycans.</td>
</tr>
<tr>
<td>SimGlycan</td>
<td><a href="http://www.premierbiosoft.com/glycan/index.html">http://www.premierbiosoft.com/glycan/index.html</a></td>
<td>Predicts the structure of glycans from MS/MS data by matching spectra to a built-in database.</td>
</tr>
</tbody>
</table>
### Table 3

**Glycosylation Site Prediction Tools.**

<table>
<thead>
<tr>
<th>Database</th>
<th>Link to Prediction Tool</th>
<th>Type</th>
<th>Overview</th>
</tr>
</thead>
</table>
Table 4
Freely Available N-linked Glycopeptide Analysis Tools.

<table>
<thead>
<tr>
<th>Analysis Tool</th>
<th>Link to Automated Program</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlycoMod</td>
<td><a href="http://web.expasy.org/glycomod/">http://web.expasy.org/glycomod/</a></td>
<td>GlycoMod determines potential glycopeptide compositions, on the basis of mass, from MS data.</td>
</tr>
<tr>
<td>GlycoPep DB</td>
<td><a href="http://hexose.chem.ku.edu/glycop.htm">http://hexose.chem.ku.edu/glycop.htm</a></td>
<td>GlycoPep DB deduces possible biologically relevant glycan compositions from glycopeptide MS data using a “smart search”.</td>
</tr>
<tr>
<td>GlycoSpectrumScan</td>
<td><a href="http://www.glycospectrumscan.org">http://www.glycospectrumscan.org</a>.</td>
<td>GlycoSpectrumScan searches LC-MS data to identify glycopeptides and determine glycoform location.</td>
</tr>
<tr>
<td>GlycoWorkbench</td>
<td><a href="http://download.glycoworkbench.org/">http://download.glycoworkbench.org/</a></td>
<td>GlycoWorkbench was designed for the annotation of glycopeptide MS/MS data through fragmentation analysis and scoring of only the glycan portion.</td>
</tr>
<tr>
<td>Glyco-Peakfinder</td>
<td><a href="http://glyco-peakfinder.org/">http://glyco-peakfinder.org/</a></td>
<td>Glyco-Peakfinder is capable of <em>de novo</em> glycopeptide analysis through glycan fragmentation profiling for spectra after the peptide sequence is input by the user.</td>
</tr>
<tr>
<td>GlycoPep ID</td>
<td><a href="http://hexose.chem.ku.edu/predictiontable2.php">http://hexose.chem.ku.edu/predictiontable2.php</a></td>
<td>GlycoPep ID works to analyze glycopeptides from MS/MS spectra of complex mixtures by identifying the peptide portion based on expected product ions.</td>
</tr>
<tr>
<td>GlycoMiner</td>
<td><a href="http://www.chemres.hu/ms/glycominer/tutorial.html">http://www.chemres.hu/ms/glycominer/tutorial.html</a></td>
<td>GlycoMiner is designed to identify glycopeptides from qTOF MS/MS data, and assigns composition for spectra of quality containing specific marker ions.</td>
</tr>
<tr>
<td>GPS</td>
<td><a href="http://edwardslab.bmcb.georgetown.edu/software/GlycoPeptideSearch.html">http://edwardslab.bmcb.georgetown.edu/software/GlycoPeptideSearch.html</a></td>
<td>GPS utilizes a glycan database to generate glycopeptide compositions after searching LC-MS/MS data of purified proteins and matching fragmentation patterns.</td>
</tr>
<tr>
<td>GlypID</td>
<td><a href="http://www.cbs.dtu.dk/services/DictyOGlyc/">http://www.cbs.dtu.dk/services/DictyOGlyc/</a></td>
<td>GlypID is designed to identify glycopeptides from LC-MS/MS experiments using a combination of MS1 and MS2 data.</td>
</tr>
<tr>
<td>GlypID 2.0</td>
<td><a href="http://mendel.informatics.indiana.edu/~chuyu/glypID/software.html">http://mendel.informatics.indiana.edu/~chuyu/glypID/software.html</a></td>
<td>GlypID 2.0 uses CID and HCD information to identify glycan type, monosaccharide composition, and attachment site for N-linked glycopeptide MS/MS data.</td>
</tr>
<tr>
<td>GPG</td>
<td><a href="http://glycopro.chem.ku.edu/GPGHome.php">http://glycopro.chem.ku.edu/GPGHome.php</a></td>
<td>GPG scores glycopeptide candidates after searching MS/MS data for the predicted product ions for each composition tested.</td>
</tr>
</tbody>
</table>