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

Carrie L. Woodin, **David Hua**, **Morgan Maxon**, **Kathryn R. Rebecchi**, **Eden P. Go**, and **Heather Desaire**[*]
Department of Chemistry, University of Kansas, Lawrence, KS

Abstract

GlycoPep Grader (GPG) is a freely-available software tool designed to accelerate the process of accurately determining glycopeptide composition from tandem mass spectrometric data. GPG relies on the identification of unique dissociation patterns shown for high mannose, hybrid, and complex *N*-linked glycoprotein types, including patterns specific to those structures containing fucose or sialic acid residues. The novel GPG scoring algorithm scores potential candidate compositions of the same nominal mass against MS/MS data through evaluation of the Y_1 ion and other peptide-containing product ions, across multiple charge states, when applicable. In addition to evaluating the peptide portions of a given glycopeptide, the GPG algorithm predicts and scores product ions that result from unique neutral losses of terminal glycans. GPG has been applied to a variety of glycoproteins, including RNase B, asialofetuin and transferrin, and the HIV envelope glycoprotein, CON-S gp140 CFI. The GPG software is implemented predominantly in PostgreSQL, with PHP as the presentation tier, and is publically accessible online. Thus far, the algorithm has identified the correct compositional assignment from multiple candidate *N*-glycopeptides in all tests performed.

INTRODUCTION

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

Typically, glycosylation analysis using mass spectrometry (MS) techniques is accomplished using one of two approaches: Glycan analysis and glycopeptide analysis. [12] The most information-rich of these methods is glycopeptide analysis, as glycosylation characteristics at individual sites of glycan attachment are readily identifiable. [2, 12] High resolution MS data is used to determine potential candidate compositions for mass spectral peaks that are suspected or known to be from glycopeptides. Computer-based programs such as Glycomod [13] and GlycoPep DB [14] calculate glycopeptide candidate compositions on the basis of mass information, as do a number of custom-generated databases. [15, 16] Unfortunately, a large amount of mass redundancy is typically encountered in glycopeptide analysis. Many different combinations of glycan composition + peptide composition are

[*]To whom correspondence should be addressed. University of Kansas, 2030 Becker Drive, Lawrence, KS 66047. Phone: 785-864-3015 hdesaire@ku.edu.

isobaric, [15] so multiple candidate compositions frequently correspond to the same nominal mass. Therefore, while high resolution MS data is useful for predicting possible glycopeptide candidate compositions, it alone is not sufficient to identify glycopeptides unambiguously. As a result, MS/MS experiments are often necessary to correctly assign glycopeptide compositions. When the analyses of these data are performed manually, the process is laborious, time-consuming, and requires significant expertise. [3, 4, 17]

A few unique strategies have been developed to automate the process of scoring MS/MS data against potential glycopeptide compositions. These include programs described in references 18, 19, 20, 21, 22. However, none of these analysis tools are freely-accessible to the public. [18, 19, 20, 21, 22] In terms of those tools that are publically available for glycopeptide analysis, many have been designed to predominantly analyze the fragmentation of glycans.[23, 24] Although these tools are capable of analyzing glycopeptides, the peptide component must be known in advance, which severely limits their utility for analysis of unknown glycopeptides. [23, 24] GlycoWorkBench, [23] and Glyco-Peakfinder [24] both utilize this approach for the annotation of glycans in glycopeptide data. A completely different approach is utilized by GlycoPep ID, a web-based tool developed by Go *et al.* [25] GlycoPep ID interprets MS/MS data of glycopeptides to identify the peptide component of glycopeptides through analysis of expected product ions, but the key disadvantage of this program is that it does not include a scoring function. [25]

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

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

EXPERIMENTAL/METHODS

Details regarding the materials and reagents, along with the experimental protocols for sample preparation and MS analysis of glycopeptides can be found in the Supplemental Materials.

Development of a Gycopeptide Training Set

In order to develop the GPG algorithm, a set of "known" glycopeptides and their MS/MS data were required; the training set included glycopeptides from RNase B, asialofetuin and transferrin, as these are well characterized samples. [27, 28, 29] To identify the glycopeptides from these samples in the MS data, a prediction table of theoretical *m/z* values corresponding to glycopeptide compositions for each of the three proteins was prepared. The amino acid sequences from RNase B, asialofetuin, and transferrin were obtained from Uniprot (www.uniprot.org) and their sequences were imported into Protein Prospector (http://prospector.ucsf.edu/prospector/mshome.htm) where tryptic peptides containing Cys residues were modified with carbamidomethylation, and a theoretical tryptic digest was performed to consider up to two tryptic miscleavages. The masses of the peptides that contained potential *N*-linked glycosylation sites were added to the masses of the known glycan compositions for each glycosylation site, in order to obtain glycopeptide masses. These masses were converted into *m/z* values corresponding to the glycopeptides in multiple charge states. The MS/MS data for RNase B, asialofetuin, and transferrin were then searched to identify spectra that corresponded to the correct *m/z* value for a given glycopeptide composition. The MS/MS data were carefully (manually) evaluated, to verify the glycopeptide assignment.

The Glycopeptide Validation Set—In order to test the GPG software, a validation set of glycopeptide compositions that were not used in the fragmentation studies or algorithm development was necessary. The *validation set* for these studies comprised data from a glycoprotein, CON-S gp140 CFI, which had been previously analyzed in our laboratory. [30] Data from this protein was selected because prior analyses demonstrated that all the necessary glycoform types were present as glycopeptides (including high mannose and complex/hybrid structures with and without sialic acid and fucose.) Additionally, since the protein has more than 25 glycosylation sites, a wide variety of glycosylated peptide sequences were also available. Furthermore, all the MS/MS data on this protein had been previously analyzed manually, as described elsewhere. [30]

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

Prior to using GlycoPep Grader, the user must first successfully complete a simple math problem embedded in a CAPTCHA (completely automated public Turing test to tell

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

Generation and Input of Glycopeptide Candidate Compositions—After the MS/ MS peak list file (along with the corresponding charge state and *m/z* of the precursor ion) is uploaded to GPG in .CSV file format, peptide compositions are input manually by listing the amino acid sequence of each glycopeptide candidate ion vertically on a separate line. The glycopeptide candidate compositions are obtained by the user through freely-accessible programs such as GlycoMod [13] or GlycoPep DB[14], or custom-generated databases.[15-16] The GPG analysis tool then quickly calculates and searches for the [peptide + core component] product ions that it predicts to be present for each of the peptide portions entered. In the next window, the glycan portions for each of the candidate glycopeptides are manually entered in the same order using the following format, where n = the number of each monosaccharide residue and Neu5Ac = sialic acid: [HexNAc]n[Hex]n[Neu5Ac]n[Fuc]n. After GPG evaluates the uploaded MS/MS peak list for product ions expected to be present for each glycan, a final score is displayed in the output for each of the user-entered glycopeptide compositions.

Generation and Input of Glycopeptide Candidate Compositions

After the MS/MS peak list file (along with the corresponding charge state and *m/z* of the precursor ion) is uploaded to GPG in .CSV file format, peptide compositions are input manually by listing the amino acid sequence of each glycopeptide candidate ion vertically on a separate line. The glycopeptide candidate compositions are obtained by the user through freely-accessible programs such as GlycoMod [13] or GlycoPep DB [14], or custom-generated databases.[15-16] The GPG analysis tool then quickly calculates and searches for the [peptide + core component] product ions that it predicts to be present for each of the peptide portions entered. In the next window, the glycan portions for each of the candidate glycopeptides are manually entered in the same order using the following format, where n = the number of each monosaccharide residue and Neu5Ac = sialic acid: [HexNAc]n[Hex]n[Neu5Ac]n[Fuc]n. After GPG evaluates the uploaded MS/MS peak list for product ions expected to be present for each glycan, a final score is displayed in the output for each of the user-entered glycopeptide compositions.

Determination of False Discovery Rates and Scoring of Candidate Compositions

Decoy candidate compositions for all data sets were generated using an in-house database where a decoy polypeptide of 50,000 amino acid residues, *Titin*, was multiplexed to a biologically relevant library of approximately 200 glycans. (These glycans are the same ones used in the on-online tool, GlycoPep DB.[14]) All selected decoy candidate compositions have a calculated neutral mass that is within 50 ppm of the FT-ICR MS monoisotopic peak value of the glycopeptide precursor ion for the CID spectrum tested. The decoy glycopeptide compositions, along with the correct glycopeptide composition assignment, were used to determine the false discovery rate of the GPG tool.

RESULTS AND DISCUSSION

GlycoPep Grader (GPG) was designed to analyze *N*-linked glycopeptides' CID data. RNase B, asialofetuin, and transferrin were chosen as model glycoproteins for the initial development of the GPG algorithm because they are well characterized and contain various glycoform types. Detailed information on the glycosylation characteristics of glycopeptides used for the development and validation of the GPG algorithm and analysis tool is included in supplementary material (Supplementary Table 1).

CID Studies and *N*-Glycopeptide Fragmentation Rules

Representative data from glycopeptides of RNase B, asialofetuin, and transferrin, are shown in Figure 1. These data show typical fragmentation patterns for glycopeptides in the following categories: A) high mannose type, B) complex or hybrid type and C) complex type structures containing the more labile residues of sialic acid and/or fucose. The CID spectra of these glycopeptides illustrate that many of the same types of product ions were detected in the glycopeptide MS/MS data, regardless of the attached glycan composition. Specifically, product ions containing the peptide and portions of the pentasaccharide core were found in all these spectra and most other spectra in the training set, regardless of the glycan type. Herein, those peptide-containing product ions are referred to as the [peptide + core component] ions. The GlycoPep Grader algorithm uses the presence of these ions to score the peptide portion of the candidate glycopeptide composition, included in Supplementary Materials. The Y_1 ion, which contains the peptide and one HexNAc residue from the pentasaccharide core, has been shown to be a highly abundant ion in MS/MS data collected on glycopeptides. This product ion is also considered a very indicative identifier of a glycopeptide's peptide portion,[31] so the GPG algorithm weights this ion more heavily and scores it on the basis of its intensity as well.

From the MS/MS data we obtained, the [peptide + core component] product ions were found to be present in multiple charge states, when the charge state of the precursor ion was greater than one, as shown in Figure 1. For all three model glycopeptides, these product ions were detected in both the precursor's charge state and the next lowest charge states, as shown by the left panel in Figure 1A-C. This finding is consistent with previous reports by Lebrilla and coworkers. [32] Therefore, the GPG algorithm scores these ions across multiple charge states. In detail, if a CID spectrum was acquired on a glycopeptide precursor ion in the 4+ charge state, the GPG software would search for *m/z* values corresponding to the [peptide + core component] product ions for the 4^+, 3^+, 2^+, and 1^+ charge state, as long as their calculated values fall within the scan range.

The second predominant type of product ion detected in the glycopeptide training set data were neutral losses of terminal monosaccharides from the glycopeptide precursor ion. In contrast to the [peptide + core component] fragmentation, neutral losses from the precursor ion observed for the three model glycoprotein types were unique to each candidate's carbohydrate composition. These ions, herein referred to as the [precursor – monosaccharide] product ions, are used to score the glycan portion of the candidate glycopeptide's composition.

An example highlighting the glycan-specific fragmentation for glycopeptides is shown in Figure 1. In the right panel of Figure 1A data for a high mannose type glycopeptide show sequential mannose losses, and the neutral loss of this residue as the predominant fragmentation for high mannose containing glycopeptides is well established. [28, 33, 34] These ions are typically present in the spectrum in the same charge state as the precursor ion.

For complex or hybrid bi- and tri-antennary structures containing no labile fucose or sialic residues, (such as the representative glycopeptide in the right panel of Figure 1B), the predominant neutral losses were found to be dependent on the total number of HexNAc vs. Hex monosaccharide residues. If there are more HexNAc residues than Hex residues, the key diagnostic loss most commonly observed in the training set was shown to be loss of two HexNAc from the glycopeptide precursor ion. In comparison, those compositions containing more Hex residues than HexNAc residues showed a key diagnostic loss corresponding to the loss of [Hex + HexNAc] from the precursor ion. Figure 1B shows an example where the [Hex + HexNAc] loss is readily detected. The GPG algorithm scores these ions to verify the glycan portion of the glycopeptides.

Finally, for CID spectra collected on glycopeptides containing labile residues such as sialic acid or fucose, the predominant [precursor – monosaccharide] product ion is the neutral loss of these labile residues from the glycopeptide precursor. For example, in the right panel of Figure 1C, loss of sialic acid is detectable as a major product ion. Often, these ions are detected in both the precursor ion's charge state, and in the charge state below that of the precursor ion. (While data for only a glycopeptide containing sialic acid is shown, glycopeptides containing at least one fucose residue generally follow the same trend, since fucose is also a more labile monosaccharide.)

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

Novel GPG Scoring Algorithm

Figure 2 summarizes a simplified version of the scoring scheme used in GPG. A more detailed version of the scoring system, as it stands currently, is available at http://glycopro.chem.ku.edu/GPGHome.php. The original scoring algorithm is also available in the Supplementary Material. For scoring the peptide portion, the [peptide + core component] product ions are calculated for the candidate glycopeptide beginning with the [naked peptide] and continuing through the [peptide + intact pentasaccharide core] for a total six possible [peptide + core component] product ions: 1. [naked peptide], 2. [peptide + HexNAc], 3. [peptide + 2HexNAc], 4. [peptide + 2HexNAc + Hex], 5. [peptide + 2HexNAc + 2Hex], and 6. [peptide + 2HexNAc + 3Hex]. Each of these ions is then searched for in the MS/MS data in multiple charge states, as explained above. The scoring algorithm for these ions does not change, regardless of the *N*-linked glycopeptide type.

Separate fragmentation rules are implemented for the glycan portion of the glycopeptide, depending on which types of glycans are present in the candidate composition. To determine which fragmentation rules apply, the glycan is first evaluated against eight possible categories including: 1) high mannose type glycans without appended fucose; 2) high mannose glycans that also contain fucose; 3) complex or hybrid structures containing sialic acid (defined as any glycan that is not in groups 1 or 2, does not contain any fucose residues, but contains sialic acid); 4) complex or hybrid type structures containing sialic acid and fucose residues (defined as any glycan that is not in groups 1 or 2, and contains both sialic acid and fucose residues); 5) complex or hybrid type structures that contain fucose and multiple terminal HexNAc residues; (defined as any glycan that is not in groups 1-4, does not contain sialic acid, and has at least one fucose residue and a greater number of HexNAc than Hex residues); 6) complex/hybrid type structures that contain fucose and terminal Hex residues (defined as any glycan that is not in groups 1-5, does not contain sialic acid, has at least one fucose residue, and has a greater number of Hex than HexNAc residues); 7)

complex/hybrid type structures with multiple terminal HexNAc residues but no sialic acid or fucose; (which is the same as group 5 glycans, except no fucose is present); and 8) complex/ hybrid type structures that lack sialic acid or fucose and contain terminal Hex residues (which is the same as group 6 glycans, except no fucose is present). The glycan classification system described above was developed to account for the fact that glycopeptides with these different glycan components fragment differently and have different diagnostic ions identifying them. This approach is also supported by recently published research that shows the types of product ions in tandem mass spectra of glycopeptides vary, depending on the unique glycan substituents present. [33] The diagnostic ions that are scored for each glycan type are depicted in Figure 2.

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

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

Candidate Composition Scoring by GPG

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

In Figure 3, an example of the candidate composition scoring by GPG is shown for a CID spectrum collected from a high mannose type glycopeptide from RNase B. The same spectrum is shown in Figures 3A, B, and C. However, each panel shows a different candidate composition for this spectrum and includes the results of how GPG scored each composition. The correct composition is in 3A, while two decoy compositions are shown in Figures 3B and 3C. The [precursor – monosaccharide] product ions searched by GPG are calculated based on the candidate composition. For candidate A, which contains a high mannose glycan, GPG predicts the sequential loss of mannose residues from the precursor ion and evaluates the [precursor – monosaccharide] product ions by searching the MS^2 peak list for the *m/z* values corresponding to sequential losses of individual hexose residues. Candidate compositions B and C are both classified as complex or hybrid glycans without sialic acid or fucose, so the same set of fragmentation rules applies for the glycan component in these two spectra. In addition to variations in the glycan scoring, each

spectrum is scored differently for the [peptide + core component] ions, because each spectrum has a different candidate peptide composition. As a result, GPG returns separate scores for the candidate compositions in B and C, even though the glycan portions are similar. The calculations for the different types of fragmentation ions are weighted by the software, with [peptide + core component] product ions accounting for 67% and [precursor – monosaccharide] product ions accounting for 33% of the score. GPG reports a final score of 97% for the correct glycopeptide assignment (candidate composition A), 20% for the first decoy glycopeptide assignment (candidate composition B) and 27% for the second decoy glycopeptide assignment (candidate composition C).

A second example of spectra scored by GPG is presented in Supplementary Figure 1. In this case, different types of [precursor – monosaccharide] product ions are searched, as different glycans are present in the candidate and decoy compositions. The CID spectrum for Supplemental Figure 1 is a sialylated glycopeptide from transferrin. Associated text in the Supplemental section also describes how GPG scores these spectra.

In Supplemental Table 2, 45 test examples are provided that show GPG scores for glycopeptides analyzed from experimental MS/MS data in the *training data set*. In each example, the correct composition is compared against at least 2 decoy compositions of the same nominal mass. A wide variety of glycopeptide compositional arrangements were tested. Over 150 glycopeptide spectra from the 45 unique glycopeptides in the training data set were scored using GPG, with the correct candidate receiving the highest score in each test performed.

GPG Validation: Application to Recombinant Gp120 HIV Envelope Glycoprotein

As the GPG algorithm was designed after studying the fragmentation patterns obtained for RNase B, asialofetuin, and transferrin, whose spectra comprise the training data set, it was expected that the automated GPG tool would perform well when testing the training data set. Therefore, after analysis of the training data set, the GlycoPep Grader software was used to analyze CID data collected on tryptic digests of the HIV envelope protein, CON-S gp140 CFI. The resulting CID spectra from the CON-S gp 140 CFI glycopeptides (herein referred to as the validation data set) contains MS/MS data on glycopeptides of varying *N*-linked glycan types and compositional arrangements. A total of over 100 CID spectra from 34 unique CON-S gp140 CFI glycopeptides were tested using the GPG tool. The results are summarized in Supplemental Table 3. A minimum of 3 candidate compositions were scored for each spectrum, with an average of 4-5 glycopeptide candidates being evaluated in each test performed. In agreement with the training data set results, the GPG algorithm assigned the highest score to each correct candidate composition, for each CON-S gp140 CFI glycopeptide spectrum, scored in the validation data set. An example of a scored fucosylated complex-type structure from CON-S gp140 CFI is shown in Figure 4. The GPG scores for decoy compositions tested against this spectrum are also reported on the spectra. This example is Test 31 of Supplementary Table 3.

While the data from both the training sets and validation sets were quite encouraging, one might note that in each case, a limited number of decoys were tested against the true composition. To test the likelihood that this limited number of decoys was a required feature for the correct candidate to get the top score, a glycopeptide spectrum from gp140 was tested against 9 alternate isobaric candidate compositions. Scores are shown in Table 1, and the MS/MS data is in Supplemental Figure 2. The correct composition of NCSFNITTEIR + [HexNAc]4[Hex]4[Fuc]1 was indicated with the highest GPG score, 80%, while the highest scoring decoy composition, INETLELLSESPVYSTK + [HexNAc]2[Hex]3[Fuc]1 was assigned a GPG score of 68%. Although the score values and distribution varies from spectrum to spectrum, GPG ranked the correct candidate composition as the most probable

glycopeptide in each test performed, including approximately 300 CID spectra from the training and validation sets. A screen shot of the GPG scoring output for a high mannose type CON-S gp140 CFI glycopeptide, along with three decoy candidate compositions, is included in Supplemental Figure 3.

CONCLUSION

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

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

REFERENCES

(1). Apweiler R, Hermjakob H, Sharon N. Biochim. Biophys. Acta. 1999; 1473:4–8. [PubMed: 10580125]

(2). Wuhrer M, Deelder AM, Hokke CH. J. Chromatogr. B. 2005; 825:124–133.

(3). Morelle W, Canis K, Chirat F, Faid V, Michalski J-C. Proteomics. 2009; 6:3993–4015. [PubMed: 16786490]

(4). Murrell MP, Yarema KJ, Levchenko A. Chem. Biochem. 2004; 5:1334–1347.

(5). Helenius A, Aebi M. Science. 2001; 291:2364–2369. [PubMed: 11269317]

(6). Petrescu A-J, Wormald MR, Dwek RA. Current Opinion in Structural Biology. 2006; 16:600–607. [PubMed: 16938454]

(7). Skropeta D. Bioorg. Med. Chem. 2009; 17:2645–2653. [PubMed: 19285412]

(8). Budnik BA, Lee RS, Steen JA. J. Biochim. Biophys. Acta. 2006; 1764:1870–1880.

(9). Bertozzi CR, Kiessling LL. Science. 2001; 291:2357–2364. [PubMed: 11269316]

(10). Zaia J. Chemistry and Biology. 2008; 15:881–892. [PubMed: 18804025]

(11). Drake PM, Cho W, Li B, Prakobphol A, Johansen E, Anderson NL, Regnier FE, Gibson BW, Fisher SJ. Clinical Chemistry. 2010; 56:223–236. [PubMed: 19959616]

(12). Dalpathado DS, Desaire H. Analyst. 2008; 133:731–738. [PubMed: 18493671]

(13). Cooper CA, Gasteiger E, Packer NH. Proteomics. 2001; 1:340–349. [PubMed: 11680880]

(14). Go EP, Rebecchi KR, Dalpathado DS, Bandu ML, Zhang Y, Desaire H. Anal. Chem. 2007; 79:1708–1713. [PubMed: 17297977]

(15). Desaire H, Hua D. Int. J. Mass. Spectrom. 2009; 287:21–26.

(16). Wang X, Emmett MR, Marshall AG. Anal. Chem. 2010; 82:6542–6548. [PubMed: 20586410]

(17). Wuhrer M, Catalina MI, Deelder AM, Hokke CH. J. Chromatogr. B. 2007; 849:115–128.

(18). Ren JM, Rejtar T, Li L, Karger BL. J. Proteome Res. 2007; 6:3162–3173. [PubMed: 17625816]

(19). Joenväärä S, Ritamo I, Peltoniemi H, Renkonen R. Glycobiology. 2008; 18:339–349. [PubMed: 18272656]

(20). Shan B, Ma B, Zhang K. J. Bioinform. Comput. Biol. 2008; 6:77–91. [PubMed: 18324747]

(21). Peltoniemi H, Joenväärä S, Renkonen R. Glycobiology. 2009; 19:707–714. [PubMed: 19270074]

(22). Goldberg D, Bern M, Parry S, Sutton-Smith M, Panico M, Morris HR, Dell A. J. Proteome Res. 2007; 6:3995–4005. [PubMed: 17727280]

(23). Ceroni A, Maass K, Geyer H, Geyer R, Dell A, Haslam SM. J. Proteome Res. 2008; 7:1650–1659. [PubMed: 18311910]

(24). Maass K, Ranzinger R, Geyer H, von der Lieth C-W, Geyer R. Proteomics. 2007; 7:4435–4444. [PubMed: 18072204]

(25). Irungu J, Go EP, Dalpathado DS, Desaire H. Anal. Chem. 2007; 79:3065–3074. [PubMed: 17348632]

(26). Ozohanics O, Krenyacz J, Ludányi K, Pollreisz F, Vékey K, Drahos L. Rapid Commun. Mass. Spectrom. 2008; 22:3245–3254. [PubMed: 18803335]

(27). Rebecchi KR, Wenke JL, Go EP, Desaire H. J. Am. Soc. Mass. Spectrom. 2009; 20:1048–1059. [PubMed: 19278867]

(28). Alley WR, Mechref Y, Novotny MV. Rapid Commun. Mass. Spectrom. 2009; 23:161–170. [PubMed: 19065542]

(29). Satomi Y, Shimonishi Y, Hase T, Takao T. Rapid Commun. Mass. Spectrom. 2004; 18:2983–2988. [PubMed: 15536627]

(30). Go EP, Irungu J, Zhang Y, Dalpathado DS, Liao H–X, Sutherland LL, Alam SM, Haynes BF, Desaire H. J. Proteome Res. 2008; 7:1660–1674. [PubMed: 18330979]

(31). Ritchie MA, Gill AC, Deery MJ, Lilley K. J. Am. Soc. Mass. Spectrom. 2002; 9:1065–1077. [PubMed: 12322954]

(32). Seipert RR, Dodds ED, Clowers BH, Beecroft SM, German JB, Lebrilla CB. Anal. Chem. 2008; 80:3684–3692. [PubMed: 18363335]

(33). Nwosu CC, Seipert RR, Strum JS, Hua SS, An HJ, Zivkovic AM, German BJ, Lebrilla CB. J. Proteome Res. 2011; 10:2612–2624. [PubMed: 21469647]

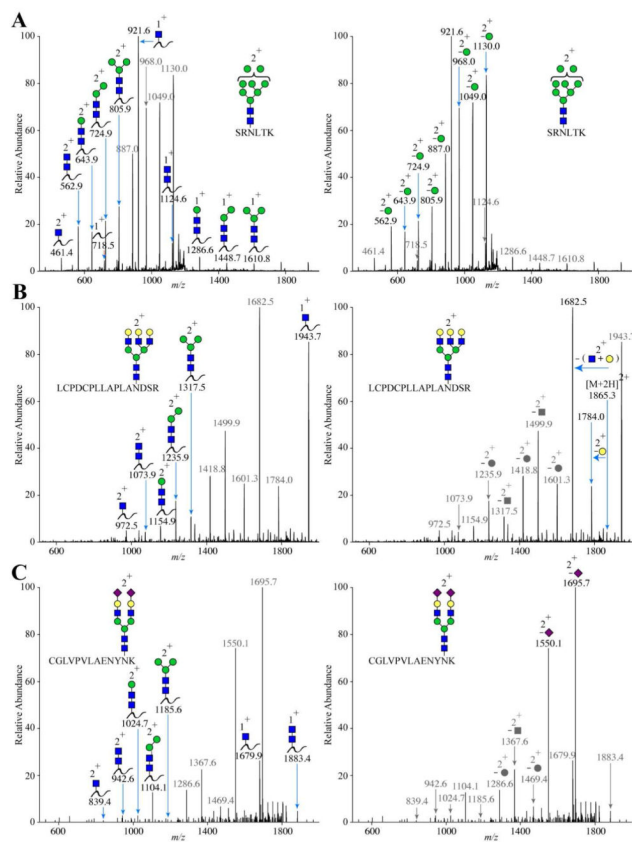(34). Zhang Z, Shah B. Anal. Chem. 2010; 82:10194–10202. [PubMed: 21090765]

Figure 1.
CID data from model *N*-linked glycopeptides used to generate GPG algorithm fragmentation rules. A high mannose glycopeptide from RNAse B is shown in A.; a sialylated complex glycopeptide from transferrin is shown in B.; and a complex glycopeptide from asialofetuin is shown in C. The spectra at left in A. – C. show the peptide-containing, or [peptide + core component], product ions detectable for all *N*-linked glycopeptides (regardless of the glycan attached), while spectra at the right in A. – C. show those product ions that result from neutral losses of monosaccharides, [precursor – monosaccharide], found to be unique to each *N*-glycan type. (Diagnostic neutral losses analyzed by GPG are shown in color, while other neutral losses that are not useful in determining the glycan type are shown in gray.)
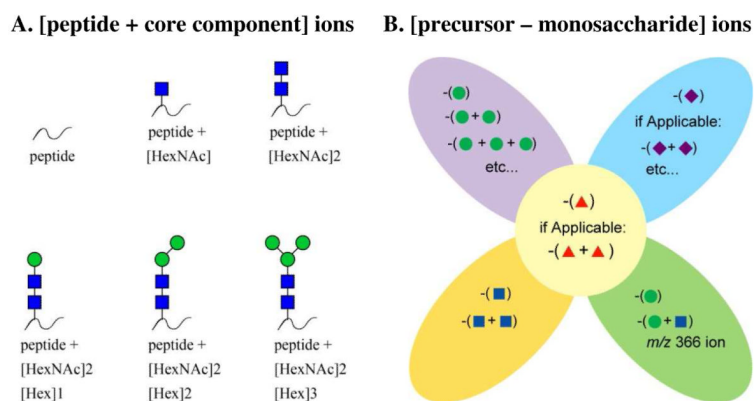
Figure 2.
Diagram of ions scored by GlycoPep Grader (GPG) algorithm for (A) [peptide + core component] and (B) [precursor – monosaccharide] product ions expected for each of the eight group types, described in the text. In (A), the six different [peptide + core component] product ions that are searched by GPG are depicted. In (B), the monosaccharide neutral losses evaluated for group 1 (see manuscript text) are shown in the purple oval; for group 2, the relevant losses are shown in both the purple oval and the yellow circle; group 3, the relevant losses are shown in the blue oval; group 4, in the blue oval and yellow circle; group 5, in the yellow oval and yellow circle; group 6, in the green oval and yellow circle; group 7 and group 8, neutral losses are in the yellow oval at bottom left, and by the green oval at bottom right, respectively.
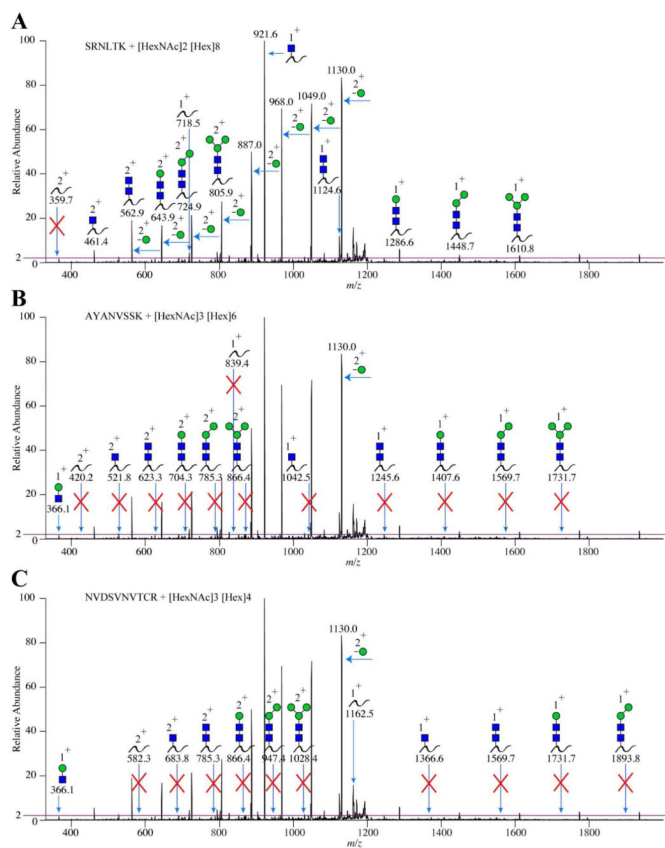
Figure 3.
CID data of an RNAse B glycopeptide from the training data. A) GPG scoring of the correct glycopeptide composition (97%) B) and C) Scoring of two decoy compositions of the same nominal mass (Scores are 20% and 27% respectively). Exact neutral masses of candidate compositions (A), (B), and (C) are, in order: 2419.9945, 2419.9733, and 2419.9978. The X's on the spectra indicate the absence of a product ion that was predicted to be present by GPG for a given candidate composition. A 2% relative abundance threshold was used for [peptide + core component] product ion matching to decrease false positives from noise.
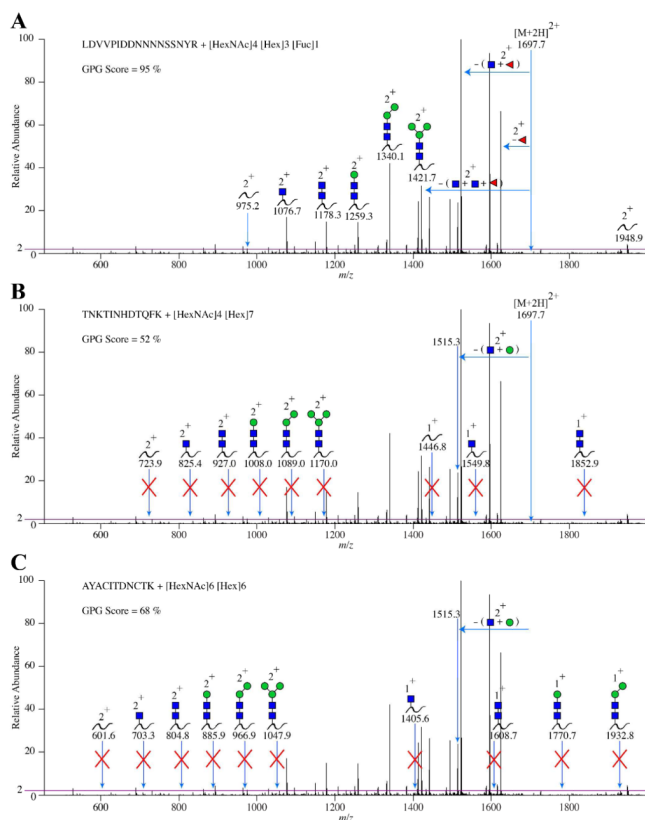
Figure 4.
MS^2 data from the validation set. (A) GPG evaluation of the correct candidate composition assignment for a fucosylated *N*-glycopeptide from CON-S gp140ΔCFI. Scoring in (B) and (C) shows evaluation of this spectrum against decoy candidate compositions with the same nominal mass. Arrows with X's indicate ions that were not present in the spectra. A 2% relative abundance threshold was used for [peptide + core component] product ion matching to decrease false positives from noise. For the composition in A, GPG generated a score of 95%. The decoy compositions in B) and C) were scored at 52% and 68% respectively, indicating that GPG scored the correct compositional assignment as the most probable composition from this pool of candidates.

Table 1

GPG score results for 10 candidate compositions tested against the MS[2] data shown in Supplementary Figure 2.

Mass	Glycopeptide Candidate [A] Composition [B]	Score [C]
2960.2338	1. NCSFNITTEIR + [HexNAc]4[Hex]4[Fuc]1	80
2960.3006	2. YILKLENSSGSK + [HexNAc]4 + [Hex]5	52
2960.3228	3. QNATVQGLIQGK + [HexNAc]6[Hex]3	39
2960.2638	4. YTVVAGGNVSTAK + [HexNAc]3[Hex]4[Neu5Ac]1[Fuc]1	11
2960.2740	5. NGTEILKSK + [HexNAc]5[Hex]5[Fuc]1	32
2960.2700	6. VENLTEGAIYYFR + [HexNAc]3[Hex]3[Neu5Ac]1	0
2960.2644	7. YTLTLENSSGTK + [HexNAc]5[Hex]3[Fuc]1	34
2960.2642	8. YILTVENSSGSK + [HexNAc]5[Hex]4	44
2960.2484	9. TKANVTVEAR + [HexNAc]3[Hex]6[Neu5Ac]1	0
2960.3632	10. INETLELLSESPVYSTK + [HexNAc]2[Hex]3[Fuc]1	68

[A] Candidate 1 is the actual glycopeptide corresponding to the MS/MS data analyzed by GPG and candidates 2-10 are decoy compositions of nearly identical neutral mass.

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

[C] Denotes GPG scores at 2% peptide normalization.