Published in final edited form as:
Anal Chem. 2013 September 3; 85(17): 8403–8411. doi:10.1021/ac401814h.

Characterizing O-linked glycopeptides by electron transfer dissociation: fragmentation rules and applications in data analysis

Zhikai Zhu, **Xiaomeng Su**, **Daniel F. Clark**, **Eden P. Go**, and **Heather Desaire***
The Ralph N. Adams Institute for Bioanalytical Chemistry and Department of Chemistry, University of Kansas, Lawrence, KS 66047, United States

Abstract

Studying protein *O*-glycosylation remains an analytical challenge. Different from *N*-linked glycans, the *O*-glycosylation site is not within a known consensus sequence. Additionally, *O*-glycans are heterogeneous with numerous potential modification sites. Electron transfer dissociation (ETD) is the method of choice in analyzing these glycopeptides since the glycan side chain is intact in ETD, and the glycosylation site can be localized on the basis of the c and z fragment ions. Nonetheless, new software is necessary for interpreting *O*-glycopeptide ETD spectra in order to expedite the analysis workflow. To address the urgent need, we studied the fragmentation of *O*-glycopeptides in ETD and found useful rules that facilitate their identification. By implementing the rules into an algorithm to score potential assignments against ETD-MS/MS data, we applied the method to glycopeptides generated from various *O*-glycosylated proteins including mucin, erythropoietin, fetuin and an HIV envelope protein, 1086.C gp120. The site-specific *O*-glycopeptide composition was correctly assigned in every case, proving the merits of our method in analyzing glycopeptide ETD data. The algorithm described herein can be easily incorporated into other automated glycomics tools.

INTRODUCTION

O-linked glycosylation typically occurs on serine and threonine residues in a protein, with the glycan portion bonded to a hydroxyl group on the amino acid's side-chain.[1] Mucin-type *O*-glycan is the most commonly seen *O*-glycosylation form; it contains an α-*N*-acetylgalactosamine (GalNAc) core structure.[2-3] Recent studies indicate that aberrant mucin-type *O*-glycosylation on membrane proteins of tumor cells is closely related to cancer metastasis, and the tumor-specific glycosylation can be mimicked to develop glycoprotein-based cancer vaccine.[4-5] It is thus a prerequisite to unravel the glycosylation profile on proteins.

However, *O*-linked glycopeptide analysis by mass spectrometry (MS) has long been a tedious task.[2, 6] Unlike *N*-linked glycosylation, no consensus sequence is available to predict potential *O*-glycosylation sites; there are eight different basic structures in mucin-type GalNAc-linked glycans with even more branching and elongations.[3, 7] In a typical analysis workflow, high resolution MS data helps to limit the number of possible candidate compositions, and MS/MS data are essential for determining the correct glycopeptide assignment.[8-9] Unfortunately, collision-induced dissociation (CID), a readily available fragmentation method, has the disadvantage that it favors carbohydrate dissociation over

*Corresponding author. Address: 2030 Becker Drive, Lawrence, KS 66047. Phone: 785-864-3015, Fax: 785-864-5396, hdesaire@ku.edu. .

peptide fragmentation in characterizing *O*-glycopeptides. While CID data are useful for inferring the glycan composition, typically one cannot be able to identify the peptide sequence or the glycosylation site using these data.[10-11]

Among tools that can generate glycopeptide backbone cleavages including higher energy collisional dissociation (HCD)[12-14], infrared multiphoton dissociation (IRMPD)[9, 15-18] and electron capture dissociation (ECD)[17-20], electron transfer dissociation (ETD)[21-25] is the most widely used one and is highly orthogonal to CID in analyzing glycopeptides. Several groups have employed ETD in large-scale glycopeptide sequencing and *O*-glycosylation site identification at the proteome level.[26-28]

While a number of applications are now published using ETD in glycoproteomics, automated analysis for *O*-glycopeptide data is virtually nonexistent. The only reported work for automated assignment of *O*-linked glycopeptides involved the usage of Protein Prospector software to identify glycopeptides bearing *O*-glycan structures of $SA_{1-0}Hex_{1-0}HexNAc$ by scoring the CID and ETD spectra.[10, 29] Nevertheless, the algorithm was developed to weight different fragment ion types based on the statistics of peptide fragmentation rules rather than glycopeptide fragmentation.[30-31] Moreover, only *O*-glycopeptides with simple glycan compositions could be searched and assigned.

In view of the urgent need to speed data interpretation in glycopeptide analysis, we report characteristic fragmentation patterns of intact *O*-glycopeptides in ETD and provide an algorithm to score ETD spectra of *O*-linked glycopeptides. Specifically, we show that the dominant fragment ion type in the glycopeptide sequence may vary with different precursor ions. In addition, for *O*-glycopeptide species at 3+ or higher charge state, doubly charged c^{2+}- and z^{2+}-ion series were frequently recorded in the high *m/z* half of the spectrum, while their singly charged counterparts were beyond the spectral mass range. These key features were incorporated into the design of an algorithm that is specifically optimized for scoring potential *O*-glycopeptide candidates against the ETD data. Using our method, site-specific assignment of *O*-glycopeptide compositions could be made a highly accurate way. The algorithm presented here may be readily incorporated into other database search engines and data analysis tools for analyzing ETD-MS/MS spectra of *O*-glycopeptides.

EXPERIMENTAL SECTION

Samples and Reagents

Bovine fetuin was purchased from Sigma Aldrich (St. Louis, MO). The HIV envelope glycoprotein, 1086.C gp120, was expressed in transiently transfected 293T cells and purified by the Duke Human Vaccine Research Institute (Durham, NC).[32] *O*-linked glycopeptides from erythropoietin and mucin-5AC were obtained from Anaspec (Fremont, CA) for direct-infusion MS experiments. Glycerol-free peptidyl-N-glycosidase F (PNGase F) cloned from *Flavobacterium meningosepticum* was purchased from New England BioLabs (Ipswich, MA). Sequencing grade trypsin was supplied by Promega (Madison, WI). Chemical reagents were of analytical purity or better.

Sample Preparation

Glycoprotein samples of 100 μg were prepared in 100 mM Tris buffer at pH 8. To remove *N*-glycans from the glycoprotein, samples were incubated with 2 μL of PNGase F (5000 units/mL) solution at 37 °C overnight. Subsequently, glycoproteins were denatured by 6 M urea and were treated with 5 mM tris(2-carboxyethyl)-phosphine (TCEP) to reduce the disulfide bonds. Following reduction, samples were alkylated with 20 mM iodoacetamide (IAM) at room temperature for 1 h in the dark. Excess IAM was quenched by adding 10 mM

dithiothreitol (DTT). Before digestion, samples were diluted to decrease the urea concentration to 1 M. Trypsin was then added to samples at a 1:30 enzyme-to-protein ratio and the digestion lasted for 18 h at 37 °C. One microliter of formic acid was added to terminate the reaction and samples were stored at −20 °C until further analysis.

LC-MS/MS

Digested glycoprotein samples were subjected to online LC-MS/MS experiments. Sample was injected onto a Vydac Capillary C_8 column (320 μm i.d. × 10 cm, 300 Å, Micro-Tech, Vista, CA) coupled to a Thermo Scientific LTQ Velos ion trap mass spectrometer (San Jose, CA) through a Waters Acquity UltraPerformance UPLC system (Milford, MA). Mobile phases consisted of solvent A: 99.9% H_2O + 0.1% formic acid and solvent B: 99.9% CH_3CN + 0.1% formic acid. The flow rate was set at 7 μL/min. A separation gradient was employed as follows: 5% solvent B for 5 min, followed by a linear increase to 40% B in 45 min, and then a ramp to 95% B in 10 min. The column was held at 95% B for 10 min and finally reequilibrated in 5% B for another 15 min. For MS settings, the ESI source had a source voltage of 2.8 kV and the capillary temperature was 250 °C. Data were obtained in the positive ion mode. One sample was analyzed in two runs that were set for CID and ETD experiments, respectively. Following a full MS scan (*m/z* 500-2000) in the enhanced scan mode, five most intense ions from the survey scan were sequentially isolated and fragmented by CID or ETD in a data-dependent fashion. The normalized collision energy was set at 35% for CID with activation time of 10 ms. The ion-ion reaction time was 90-150 ms for ETD, and supplemental activation was turned on.[33] The automatic gain control (AGC) target value was set at 2×10^4 for the MS/MS experiment in the linear ion trap, and the AGC target value was 2×10^5 for the fluoranthene reagent anions.

Direct-infusion MS/MS

Glycopeptide standards having sequences of GT***T***PSPVPTTSTTSAP, GTTPSPVPTTST***T***SAP and EAISPPDAA***S***AAPLR (where ***T*** and ***S*** are residues modified with *N*-acetylgalactosamine, GalNAc), respectively, were dissolved in water/methanol (50:50) with 1% formic acid to a concentration of 500 nM. The prepared solution was introduced into the mass spectrometer by direct infusion at a flow rate of 3 μL/min in the positive ion mode. The ESI source was optimized using the following conditions: the spray voltage was 3.0 kV, capillary temperature was 200 °C and nitrogen carrier gas was 10 psi. Selected precursor ions in the full MS scan were subjected to both CID and ETD with a 2.5 Da isolation width. The activation time was 30 ms and activation energy was 30% in CID, while the reaction time in ETD was 100 ms with the maximum injection time of fluoranthene anions set as 150 ms. Thirty scans, each with 10 microscans, were averaged in the collection of MS/MS data.

Data Analysis

Glycoproteins used in this study (fetuin and HIV Env glycoprotein) have been well characterized in the literature regarding their *O*-glycosylation profiles and *O*-glycopeptides with known structures were searched for in the MS and MS/MS data.[32, 34-35] Specifically, glycoproteins were tryptically digested *in silico* to produce peptides with up to 2 missed cleavages. Cysteine residues were carbamidomethylated. Theoretical masses of potential *O*-glycopeptides were calculated by adding site-specific *O*-glycan masses to the corresponding peptide sequences that contain the reported glycosylation sites. Glycopeptide masses were then converted to theoretical *m/z* values, which were searched against the full scan mass spectra with a mass tolerance of 200 ppm. If a peak was matched, the CID-MS/MS spectrum was interrogated to confirm the presence of oxonium ions [m/z 204 (HexNAc), 292 (SA), 366 (Hex-HexNAc), etc.] and characteristic peaks derived from monosaccharide losses. If

the CID data was confirmed to be from a glycopeptide, the ETD-MS/MS spectrum of the same *O*-glycopeptide was verified manually to analyze its fragmentation patterns in ETD.

Algorithm Performance Test

An in-house program, GlycoPep Scorer, was coded in MATLAB based on the algorithm that was described below and in Supporting Information. A peak list file was created from each glycopeptide ETD-MS/MS spectrum and was uploaded to both GlycoPep Scorer and Protein Prospector (http://prospector.ucsf.edu) for scoring.[36] The *m/z* value and charge state of the precursor ion were input into each program. The glycoprotein sequence and randomized decoy sequences were directly entered into Protein Prospector. The same glycopeptide candidates were also scored by GlycoPep Scorer. Search parameters were set the same for the two programs. Trypsin was selected as the enzyme and 2 missed cleavage sites were set as the maximum. Carbamidomethylation was a fixed modification of cysteines. Mass accuracy was set to 20 ppm for precursor ions and 1.0 Da for fragment ions. In GlycoPep Scorer, the *O*-glycan composition (in the form of $[SA]_n[Hex]_n[HexNAc]_n$) and glycosylation site were entered for each candidate. In Protein Prospector, *O*-glycans were set as variable modifications on Ser and Thr residues in the form of HexNAc, $Hex_1HexNAc$, $SA_1HexNAc$, $SA_1Hex_1HexNAc$ or $SA_2Hex_1HexNAc$. All glycopeptide identifications were manually inspected to determine the false-discovery rates.

RESULTS AND DISCUSSION

Fetuin and the HIV Env glycoprotein, 1086.C gp120, are both *N*- and *O*-glycosylated.[32, 34-35] As a result, PNGase F was used to cleave the *N*-glycans off the proteins prior to tryptic digestion. In this way *N*-linked glycopeptides would not interfere with the analysis of *O*-linked glycopeptide data.[20] For each *O*-glycopeptide studied, its sequence, glycan composition and attachment site were confirmed based on prior knowledge of the protein, combined with the MS and CID/ETD-MS/MS data. The ETD data were specifically studied to discover distinct fragmentation patterns and to develop rules that can aid in the identification of *O*-glycopeptides.

O-Glycopeptide Fragmentation Rules in ETD

O-glycopeptide ions with *m/z* values over 1200 generally did not produce enough peptide fragments that could be used for sequencing. Below this value, c- and z-ions were frequently recorded in ETD spectra for glycopeptides of 2+ and higher charge states, along with y-ions and occasional peaks from glycan dissociations. However, the dominant fragment ion series varied significantly for different precursor ions, and even *O*-glycopeptides with analogous structures had distinct dissociation patterns. Figure 1A and 1B show the ETD-MS/MS data of two isomeric glycopeptides from mucin that only differ in their *O*-glycosylation sites. For the glycopeptide whose glycan attaches to Thr-3, c-ion series (c_8-c_{14}) are predominantly present in its ETD spectrum while no z-ions are found (Figure 1A). This pattern contrasts with the data from the Thr-13 glycosylated isomer, which generated both c- and z-ions during ETD (Figure 1B). Even for the same glycopeptide species, the ETD fragmentation may be drastically different if the charge state changes. The ETD spectra of an erythropoietin *O*-glycopeptide with 2+ and 3+ charges are demonstrated in Figure 1C and 1D, respectively. The doubly charged precursor ion primarily dissociates into eight dominant z-ions with only one single c-ion (c_{14}) produced in the spectrum (Figure 1C). As the glycopeptide carries more charges, however, its fragmentation efficiency in ETD improves so that both c- and z-ion series of high sequence coverage are recorded (Figure 1D). An effective algorithm for scoring *O*-glycopeptide ETD data must be optimized to score these types of spectra, where the fragment ion series is varied and unpredictable. Therefore, fixed weightings for different ion types would most likely not work optimally.

Furthermore, we discovered that for *O*-glycopeptides at 3+ or higher charge state, doubly charged fragment ions were likely to appear in the high *m/z* end of the ETD spectra. Example data are shown in Figure 2A and 2B, in which the precursor ions are two glycopeptides from the HIV envelope glycoprotein, 1086.C gp120, with 3+ and 4+ charges. For the glycopeptide in Figure 2A, the relatively large *O*-glycan modification (+1312.5 Da) makes the c-ions (c_{12}-c_{14}) and z-ions (z_7-z_{14}) that contain the glycosylated Thr-12 too large to be detected in the scan range of up to *m/z* 2000. Consequently, searching for these singly charged fragment ions is not very useful for increasing the coverage of the glycopeptide sequence, especially for z-ion series among which only two ions (z_3 and z_6) are found in the spectrum. By considering doubly charged c^{2+}- and z^{2+}-ions whose singly charged counterpart ions are beyond the mass range, the coverage of both c- and z-ion series are increased, as two more c^{2+}-ions and five z^{2+}-ions are identified as shown in Figure 2A. The same trend is observed in Figure 2B, where singly charged c-ions beyond c_{11} and z-ions beyond z_{10} are not recorded due to their high *m/z* values. However, the extra five c^{2+}-ions (c_{12}^{2+}-c_{16}^{2+}) and z^{2+}-ions (z_{10}^{2+} and z_{13}^{2+}-z_{16}^{2+}) provide extended sequence coverage for the *O*-glycopeptide. The fragmentation of *O*-glycopeptides in ETD also differs from *N*-linked glycopeptides significantly. As is illustrated in the ETD spectrum of a complex-type *N*-glycopeptide from avidin (Figure 2C), only singly charged c- and z-ion series exist while no c^{2+}- or z^{2+}-ions are generated except a single z_{16}^{2+} ion. In this circumstance, incorporating doubly charged fragment ions into the search of c- and z-ions is not helpful for identifying the correct glycopeptide composition, since it can lower the percentage of matched fragment ions over the number of possible ions being searched, and the false positive identifications would be increased.

An ETD spectrum of a mucin-type core-1 *O*-glycopeptide is present in Figure 3A. The most significant spectral feature is that the major peaks in the ETD spectrum are unreacted precursor ion, charge-reduced species and their neutral losses, which are not useful for identifying the glycopeptide sequence. However, the peptide backbone fragment ions (c- and z-ions) do exist in the data, as is illustrated in the two enlarged windows in Figure 3A, even though their relative intensities are very low compared to the base peak. Moreover, by comparing the two insets (*m/z* 300-500 *v.s. m/z* 1150-1350) in Figure 3A, it is found that interfering peaks are not evenly populated along the *m/z* scale in the ETD spectrum. The low *m/z* area has fewer peaks of noise even though the spectral intensity is low (normalized level of 2.35×10^2), while abundant interfering peaks are present in the high *m/z* end with relatively high intensity values (normalized level of 3.49×10^3). The trend is similar to what has been observed for *N*-linked glycopeptides, where useful fragment ions need to be differentiated from ETD side products and noise peaks.[37]

Algorithm Design and Implementation in *O*-glycopeptide ETD Data Analysis

After the key features of *O*-glycopeptide fragmentation in ETD were identified, an algorithm was developed based on the characteristic fragmentation rules. First, we employed the spectral preprocessing approach that was previously used for handling *N*-glycopeptide data to filter noise peaks in the *O*-linked glycopeptide ETD spectra.[37] Briefly, the precursor ion, charge-reduced species and their neutral losses are removed. Subsequently, the spectrum is split into two halves by the precursor *m/z* value: For the low *m/z* half, the 5 highest peaks in every 100 Da bin are retained with other peaks removed; for the high *m/z* half, only the top 3 peaks are preserved in each bin. Finally, the retained peaks in the low *m/z* area are amplified by a factor of 5. By this method, the fragment ion peaks of low *m/z* values and low intensity (but good signal-to-noise ratio), as opposed to high noise peaks in the high *m/z* area, can be preserved and given more weighting in the scoring process.

After spectral filtering, the spectrum is then subjected to algorithm scoring. As is discussed earlier, for *O*-glycopeptides, different types of peptide backbone fragment ions have large deviations of sequence coverage in ETD, and one ion series that dominates a spectrum may not be well represented in the other spectrum. As a result, in our designed algorithm, different fragment ion series (c-, z- and y-ions) are separately searched and scored. In addition, for *O*-glycopeptides at 3+ or higher charge state, doubly charged c^{2+}- and z^{2+}-ions, of which the equivalent singly charged ions are beyond the scan range, are also incorporated into the search of c- and z-ion series, thus taking advantage of the distinct *O*-glycopeptide fragmentation pattern in ETD that doubly charged ions are frequently found in the high *m/z* end of a spectrum. It should be noted that although programs used for analyzing peptide MS/MS data also consider doubly charged fragment ions,[38-40] our algorithm differs in that no doubly charged ion present in the low *m/z* half of the spectrum is searched, because these ions are typically not seen in the *O*-glycopeptide ETD data. Therefore, an individual score corresponding to each ion type, is determined by the probability that a random sequence would have the same or higher number of matched peaks as the input glycopeptide candidate, using the following equation:

$$\text{Score}\,(k-\text{ion}) = -10 \times \log\left[\sum_{k=n}^{N}\begin{pmatrix} N \\ k \end{pmatrix} p^k (1-p)^{N-k}\right]$$

Herein, *N* is the total number of searched *k*-ions, and *n* is the number of matched *k*-ions to the spectrum. In the next step, the weighting of each ion series is calculated by dividing the intensities of spectral peaks matched to the specific ion type into the total intensities of all matched peaks, and the total score of the candidate is then determined by summing up the weighted individual scores of c-, z- and y-ion series:

$$\text{Total}\ \ \text{Score} = \sum_{k=c,z,y}\left[\frac{\Sigma\text{Int.}\,(k-\text{ions})}{\Sigma\text{Int.}\,(\text{all ions})} \times \text{Score}\,(k-\text{ion})\right]$$

A detailed description of the algorithm, including the spectral filtering and scoring method, is contained in Supporting Information.

Algorithm Scoring of *O*-glycopeptide Candidate Compositions

As an example of using the algorithm in *O*-glycopeptide data analysis, the raw ETD spectrum, as illustrated in Figure 3A, was scored against the correct glycopeptide candidate of VVEIKPLGVAP***T***EAK (where ***T*** is modified with $SA_2Hex_1HexNAc$). After spectral filtering, the processed spectrum is shown in Figure 3B, where peaks that are matched to predicted fragment ions are labeled in color. The scoring parameters of all the ion series (c-, z- and y-ions) are also listed in the inset of Figure 3B. For the correct candidate, 9 singly charged c-ions and 3 doubly charged c^{2+}-ions (starting from c_{12}^{2+} of *m/z* 1085) are searched against the processed spectrum, and 7 out of the 12 c-ions searched are matched to the ETD data. Consequently, for c-ion series, the probability that a random glycopeptide sequence has seven or more c-ions matched in the spectrum, is calculated using the binomial distribution. An individual score of 54.2 for c-ions is then determined by converting the probability into the Log10 scale and multiplying by −10. Scores of z-ions and y-ions are computed in the same way, except that for y-ions, no doubly charged species are considered because they are not consistently produced. Subsequently, each ion series is weighted to calculate the total score of the input candidate, and the weighting factor is proportional to the matched spectral peaks' intensities, as described in detail in the algorithm in Supporting Information. As is shown in Figure 3B, multiple peaks assigned to c- and z-ions are dominant peaks in the

processed spectrum, and large weightings of 54% and 45%, respectively, are given to these two ion types automatically. In contrast, the y-ion series is only 1% weighted, since only two y-ions are matched to spectral peaks of low intensity. A total score of 39.9 is then determined by summing up the product of the individual ion series' score and the weighting. Clearly, we designed the algorithm with these novel weighting features because the weighting for respective ion series varies according to the assigned spectral peaks, which is ideal for the *O*-linked glycopeptide ETD data shown here. Even when one type of ions is seriously underrepresented, that ion series' score will not have a high impact in the total score because its overall intensity level is low.

The correct glycopeptide composition, as illustrated in Figure 3B, was further corroborated by the corresponding CID spectrum, which is shown in Figure 3C. The dominant peaks in the CID data are intact peptides with sequential losses of monosaccharide units, and the glycan portion can be deduced to be $SA_2Hex_1HexNAc$ based on these fragment ions (Figure 3C).

To test whether the algorithm is effective in differentiating the correct *O*-glycopeptide composition from multiple decoy candidates, the ETD data presented in Figure 3A, was further scored against nine isobaric decoy compositions bearing identical or similar *O*-glycan portions, and the result is summarized in Table 1. One can clearly see that the correct glycopeptide composition received the highest score of 39.9, which is significantly higher than other decoys. Among all the incorrect assignments, the glycopeptide candidate having the sequence YKVVEIKPLGVAP***T***EAK (where ***T*** is attached to $SA_1Hex_1HexNAc$), received the best score of 13.8. The slightly higher score for this candidate is expected, because its sequence is highly homologous to the correct glycopeptide sequence. In this case, the algorithm still works effectively to distinguish the true candidate from the incorrect composition based on the ETD data even if their sequences are very similar.

Analysis of *O*-linked Glycopeptide ETD Data Sets by GlycoPep Scorer

We integrated the spectral preprocessing method and the designed scoring algorithm into a standalone program, GlycoPep Scorer, and used the software to analyze the collected *O*-glycopeptide ETD-MS/MS data from multiple glycoproteins including mucin, fetuin, erythropoietin and the HIV envelope protein, 1086.C gp120. More than 40 ETD spectra from 22 distinct *O*-glycopeptides were scored by the program, including 5 *O*-glycopeptide species that have more than one potential glycosylation site of Ser or Thr. For every tested ETD spectrum, site-specific assignment of the corresponding *O*-glycopeptide composition was made correctly by GlycoPep Scorer, and the real glycopeptide was assigned a score at least 1.5 times higher than other decoy candidates, including positional isomers where the same peptide sequence and glycan portion are present, but the decoys differ from the correct candidate only in the glycosylation site location. The scoring results of all glycopeptide candidates using GlycoPep Scorer are summarized in Table S1 in Supporting Information.

To compare the performance of GlycoPep Scorer with other software, the same *O*-linked glycopeptide ETD data sets were also analyzed by Protein Prospector. Since the possible glycan modification is limited to simple *O*-glycan compositions, a subset of ETD spectra collected from 16 distinct glycopeptides having glycan compositions of $SA_{0\text{-}2}Hex_{0\text{-}1}HexNAc$, were subjected to Protein Prospector scoring. Table 2 lists the comparison of the results from the two programs. For the 16 unique *O*-glycopeptide spectra analyzed by Protein Prospector, 3 glycopeptide compositions were incorrectly assigned. In contrast, no decoy glycopeptide composition received a higher score than the correct glycopeptide candidate in GlycoPep Scorer, both for the subset data where the *O*-glycan conforms to the composition of $SA_{0\text{-}2}Hex_{0\text{-}1}HexNAc$, and for the whole data set, in which the glycan has a composition of $SA_{0\text{-}2}Hex_{0\text{-}2}HexNAc_{1\text{-}2}$. A full list of the test results for all

glycopeptide compositions scored by GlycoPep Scorer and Protein Prospector, is provided in Table S1 and S2, respectively, in Supporting Information. The peak lists from the ETD spectra are also provided in Supporting Information. The average scores for the correct composition and the best-matched decoy candidate given by both programs are presented in Table 2. For Protein Prospector, the correct glycopeptide assignment receives an average score of 44.7, and the highest decoy score averages at 29.7. A lager score difference is observed in GlycoPep Scorer for the same subset data, in which the correct composition has a score of 55.7 while the best-matched decoy composition has a score of 18.4. Although direct comparison of the absolute score values would be inappropriate since the two software's algorithms are different, GlycoPep Scorer is demonstrated herein to be efficacious in assigning site-specific *O*-glycopeptides accurately by analyzing the ETD data. Additionally, the larger score difference between the correct and incorrect assignment in GlycoPep Scorer provides more confidence that the highest scoring candidate is the right glycopeptide composition. The superior performance of the program in turn proves the advantage of using our spectral filtering approach and a scoring algorithm designed specifically for fragmentation of *O*-linked glycopeptides. At the current stage, the glycopeptide candidates need to be input manually into GlycoPep Scorer, which is probably the key drawback to using the software in its current format. However, the algorithm for scoring O-linked glycopeptides could be incorporated into any other glycopeptide scoring tool which uses a more automated workflow; in doing so, the convenience of automation could be combined with the power of a highly tuned scoring system.

CONCLUSIONS

We studied the fragmentation of *O*-linked glycopeptides in ETD and identified their characteristic spectral features that can be applied into data analysis automation. For *O*-glycopeptides, the dominant ion series varies with different precursor ions, and it is not uncommon to see the phenomenon that one type of ion series is much more abundant than other ion types in ETD. Furthermore, we found that doubly charged c^{2+}- and z^{2+}-ions are often recorded in the high *m/z* half of the spectra for highly charged glycopeptides, and these ions can be included into the search of c- and z-ion series to substitute the singly charged c- and z-ions of which the *m/z* values are above the mass limit. In this way the sequence coverage is increased, and the individual scores of each ion series are not undermined by a lack of doubly charged ions in the low *m/z* end.

By correlating the weighting for each type of ions with the intensity of matched peaks, we developed an algorithm that uses *O*-glycopeptide fragmentation patterns to score the potential glycopeptide compositions against the ETD data. The algorithm, along with a spectral filtering method, was combined into the GlycoPep Scorer program, which was used in data analysis of *O*-glycopeptide ETD-MS/MS spectra. The program was able to determine the site-specific *O*-glycopeptide composition correctly with no false positives, and the large score differences between the true and decoy candidates demonstrate the benefit of the algorithm in interpreting glycopeptide ETD data. The fragmentation rules and algorithm in this study can be widely applied into other computer programs for identifying *O*-glycopeptides and determining the modification site on a large scale.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute of Health (grant RO1RR026061). The HIV envelope protein sample was kindly provided by Dr. Barton Haynes and Dr. Larry Liao.

REFERENCES

(1). Van den Steen P, Rudd PM, Dwek RA, Opdenakker G. Crit. Rev. Biochem. Mol. Biol. 1998; 33:151–208. [PubMed: 9673446]

(2). Zauner G, Kozak RP, Gardner RA, Fernandes DL, Deelder AM, Wuhrer M. Biol. Chem. 2012; 393:687–708. [PubMed: 22944673]

(3). Jensen PH, Kolarich D, Packer NH. Febs J. 2010; 277:81–94. [PubMed: 19919547]

(4). Tarp MA, Clausen H. Biochim. Biophys. Acta-Gen. Subj. 2008; 1780:546–563.

(5). Tian E, Ten Hagen KG. Glycoconjugate J. 2009; 26:325–334.

(6). Burlingame AL. Curr. Opin. Biotechnol. 1996; 7:4–10. [PubMed: 8742373]

(7). North SJ, Hitchen PG, Haslam SM, Dell A. Curr. Opin. Struct. Biol. 2009; 19:498–506. [PubMed: 19577919]

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

(9). Seipert RR, Dodds ED, Lebrilla CB. J. Proteome Res. 2009; 8:493–501. [PubMed: 19067536]

(10). Darula Z, Chalkley RJ, Lynn A, Baker PR, Medzihradszky KF. Amino Acids. 2011; 41:321–328. [PubMed: 20652609]

(11). Perdivara I, Petrovich R, Allinquant B, Deterding LJ, Tomer KB, Przybylski M. J. Proteome Res. 2009; 8:631–642. [PubMed: 19093876]

(12). Segu ZM, Mechref Y. Rapid Commun. Mass Spectrom. 2010; 24:1217–1225. [PubMed: 20391591]

(13). Scott NE, Parker BL, Connolly AM, Paulech J, Edwards AVG, Crossett B, Falconer L, Kolarich D, Djordjevic SP, Hojrup P, Packer NH, Larsen MR, Cordwell SJ. Mol. Cell. Proteomics. 2011; 10:1–18.

(14). Hart-Smith G, Raftery MJ. J. Am. Soc. Mass Spectrom. 2012; 23:124–140. [PubMed: 22083589]

(15). Fukui K, Takahashi K. Anal. Chem. 2012; 84:2188–2194. [PubMed: 22300132]

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

(17). Adamson JT, Hakansson K. J. Proteome Res. 2006; 5:493–501. [PubMed: 16512663]

(18). Hakansson K, Cooper HJ, Emmett MR, Costello CE, Marshall AG, Nilsson CL. Anal. Chem. 2001; 73:4530–4536. [PubMed: 11575803]

(19). Renfrow MB, Mackay CL, Chalmers MJ, Julian BA, Mestecky J, Kilian M, Poulsen K, Emmett MR, Marshall AG, Novak J. Anal. Bioanal. Chem. 2007; 389:1397–1407. [PubMed: 17712550]

(20). Halim A, Ruetschi U, Larson G, Nilsson J. J. Proteome Res. 2013; 12:573–584. [PubMed: 23234360]

(21). Wang DD, Hincapie M, Rejtar T, Karger BL. Anal. Chem. 2011; 83:2029–2037. [PubMed: 21338062]

(22). Snovida SI, Bodnar ED, Viner R, Saba J, Perreault H. Carbohydr. Res. 2010; 345:792–801. [PubMed: 20189550]

(23). Han H, Xia Y, Yang M, McLuckey SA. Anal. Chem. 2008; 80:3492–3497. [PubMed: 18396915]

(24). Thaysen-Andersen M, Wilkinson BL, Payne RJ, Packer NH. Electrophoresis. 2011; 32:3536–3545. [PubMed: 22180206]

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

(26). Chalkley RJ, Thalhammer A, Schoepfer R, Burlingame AL. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:8894–8899. [PubMed: 19458039]

(27). Steentoft C, Vakhrushev SY, Vester-Christensen MB, Schjoldager K, Kong Y, Bennett EP, Mandel U, Wandall H, Levery SB, Clausen H. Nat. Methods. 2011; 8:977–982. [PubMed: 21983924]

(28). Vakhrushev SY, Steentoft C, Vester-Christensen MB, Bennett EP, Clausen H, Levery SB. Mol. Cell. Proteomics. 2013; 12:932–944. [PubMed: 23399548]

(29). Darula Z, Chalkley RJ, Baker P, Burlingame AL, Medzihradszky KF. Eur. J. Mass Spectrom. 2010; 16:421–428.

(30). Chalkley RJ, Medzihradszky KF, Lynn AJ, Baker PR, Burlingame AL. Anal. Chem. 2010; 82:579–584. [PubMed: 20028093]

(31). Baker PR, Medzihradszky KF, Chalkley RJ. Mol. Cell. Proteomics. 2010; 9:1795–1803. [PubMed: 20513802]

(32). Go EP, Liao HX, Alam SM, Hua D, Haynes BF, Desaire H. J. Proteome Res. 2013; 12:1223–1234. [PubMed: 23339644]

(33). Swaney DL, McAlister GC, Wirtala M, Schwartz JC, Syka JEP, Coon JJ. Anal. Chem. 2007; 79:477–485. [PubMed: 17222010]

(34). Carr SA, Huddleston MJ, Bean MF. Protein Sci. 1993; 2:183–196. [PubMed: 7680267]

(35). 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]

(36). Chalkley RJ, Baker PR, Huang L, Hansen KC, Allen NP, Rexach M, Burlingame AL. Mol. Cell. Proteomics. 2005; 4:1194–1204. [PubMed: 15937296]

(37). Zhu Z, Hua D, Clark DF, Go EP, Desaire H. Anal. Chem. 2013; 85:5023–32. [PubMed: 23510108]

(38). Hogan JM, Higdon R, Kolker N, Kolker E. Omics. 2005; 9:233–250. [PubMed: 16209638]

(39). Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang XY, Shi WY, Bryant SH. J. Proteome Res. 2004; 3:958–964. [PubMed: 15473683]

(40). Deutsch EW, Shteynberg D, Lam H, Sun Z, Eng JK, Carapito C, von Haller PD, Tasman N, Mendoza L, Farrah T, Aebersold R. Proteomics. 2010; 10:1190–1195. [PubMed: 20082347]

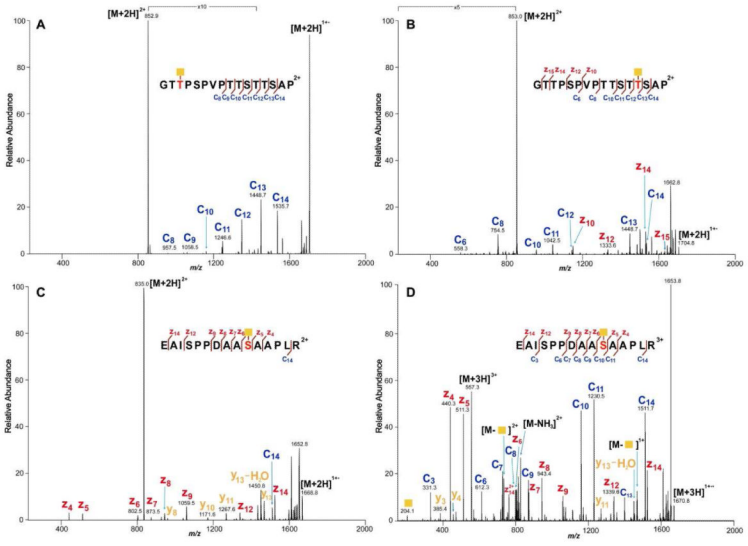

Figure 1.

ETD-MS/MS data from (A) a mucin *O*-linked glycopeptide of which the glycan is attached to the Thr-3 residue (2+, *m/z* 852.9); (B) an isomeric *O*-glycopeptide of (A) that has the same composition but with a different modification site at Thr-13 (2+, *m/z* 852.9); (C) a doubly charged *O*-glycopeptide from erythropoietin with the Ser-10 residue glycosylated (*m/z* 834.9); (D) the same glycopeptide as (C) but at 3+ charge state (*m/z* 557.0). Peptide backbone fragment ions (c-, z- and y-ions) are labeled in different colors as shown in the figure. Glycan symbols used herein and in the following figures include *N*-acetylhexosamine (yellow and blue squares, HexNAc), hexose (yellow and green circles, Hex), and sialic acid (purple diamond, SA).

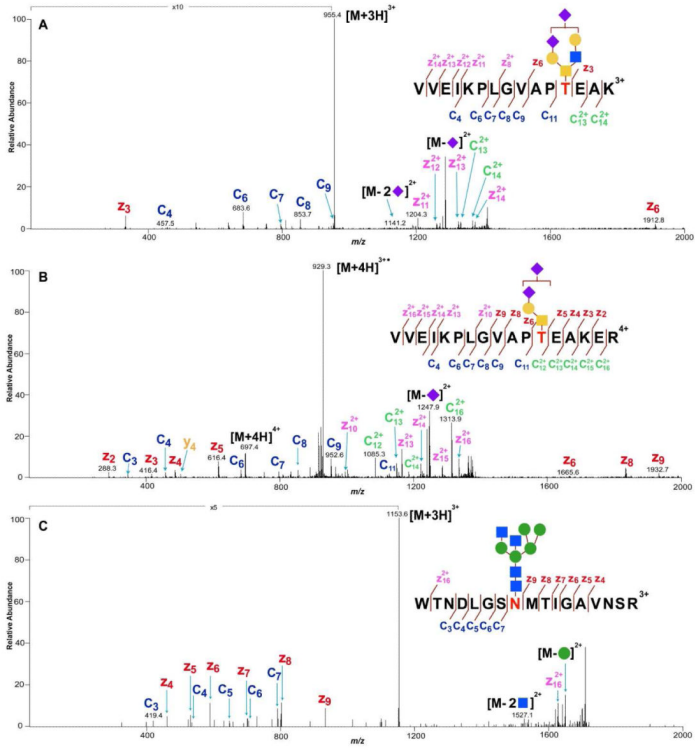

Figure 2.

ETD spectra of (A) an *O*-linked core-2-type glycopeptide (3+, *m/z* 955.1) and (B) a core-1-type *O*-glycopeptide (4+, *m/z* 696.6) from the HIV envelope glycoprotein, and (C) a hybrid-type *N*-linked glycopeptide (3+, 1153.5) from avidin.

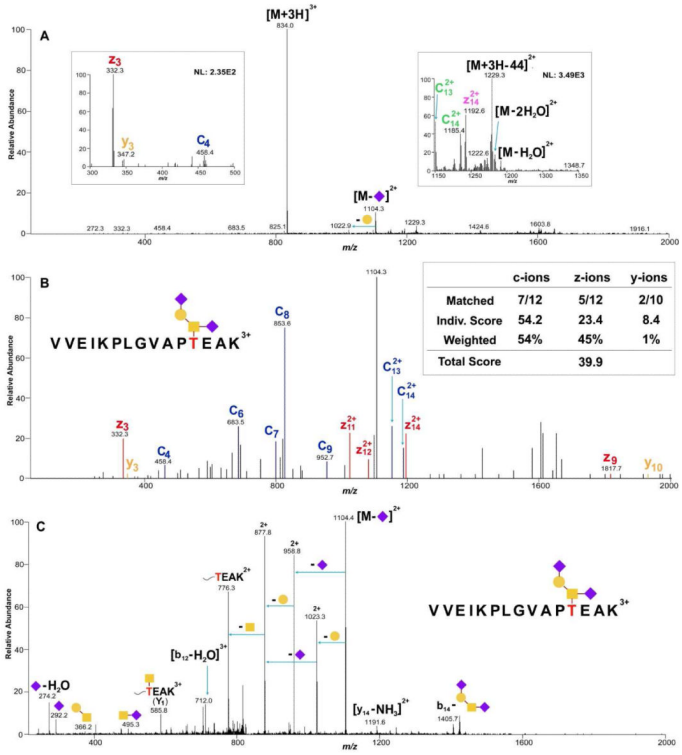

Figure 3.

(A) ETD-MS/MS data of an *O*-linked glycopeptide (3+, *m/z* 833.4) with its composition shown in (B), note that two enlarged windows showing the zoomed *m/z* regions of 300-500 and of 1150-1350, respectively are also present in the figure; (B) processed ETD data of (A) after spectral filtering to remove noise peaks, and the inset table lists the scoring results (including the individual ion series' scores and the respective weightings) of the correct glycopeptide composition against the processed data; (C) CID-MS/MS data of the same glycopeptide as shown in (B).

Table 1

Algorithm scoring results of the ETD-MS/MS data against 10 *O*-glycopeptide compositions[a]

Candidate	Mass	*O*-linked Glycopeptide Compositions[b]	Total Score
Correct	2497.2309	VVEIKPLGVAPTEAK + $SA_2Hex_1HexNAc$	39.9
Decoy 1	2497.2930	YKVVEIKPLGVAPTEAK + $SA_1Hex_1HexNAc$	13.8
Decoy 2	2497.2013	DFAGITGAYGAVAAGASFLFAR + $Hex_1HexNAc$	10.0
Decoy 3	2497.2224	YLTAPTITSGGNPPAFSLTSDGK + HexNAc	8.1
Decoy 4	2497.2057	ATIIVHLNESVNIK + $SA_2Hex_1HexNAc$	6.3
Decoy 5	2497.2309	AETPAVGLPKIEVVK + $SA_2Hex_1HexNAc$	5.7
Decoy 6	2497.2210	LAIIQFISGNPLHK + $SA_2Hex_1HexNAc$	3.9
Decoy 7	2497.2516	TLFWTAVFLTIIGFGR + $SA_1Hex_1HexNAc$	3.4
Decoy 8	2497.2356	INSLVACGENINALLIK + $SA_1Hex_1HexNAc$	3.1
Decoy 9	2497.2422	GDNLLPAIVGLSILR + $SA_2Hex_1HexNAc$	1.9

[a] The ETD spectrum is shown in Figure 3A.

[b] The *O*-glycosylation sites are labeled in red, and the monoisotopic masses of the listed glycopeptide candidates are within 20 ppm mass error.

Table 2

Analysis summary of GlycoPep Scorer and Protein Prospector in interpreting *O*-glycopeptide ETD datasets

Program Name	False Positives	Correct Score	Best Decoy Score
Protein Prospector[a] (*O*-glycan: $SA_{0-2}Gal_{0-1}GalNAc_1$)	3/16	44.7	29.7
GlycoPep Detector[a] (*O*-glycan: $SA_{0-2}Gal_{0-1}GalNAc_1$)	0/16	55.7	18.4
GlycoPep Detector[b] (*O*-glycan: $SA_{0-2}Gal_{0-2}GalNAc_{1-2}$)	0/22	53.0	18.3

[a] False positives and average scores were based on the scoring of a subset of 16 distinct ETD spectra collected from *O*-glycopeptides bearing glycan compositions of $SA_{0-2}Gal_{0-1}GalNAc_1$.

[b] False positives and average scores were based on the scoring of a total of 22 ETD spectra collected from *O*-glycopeptides bearing glycan compositions of $SA_{0-2}Gal_{0-2}GalNAc_{1-2}$.