Identification of glycopeptides with multiple hydroxylysine O-glycosylation sites by tandem mass spectrometry

Yanlin Zhang,[†] Chuan-Yih Yu,[‡] Ehwang Song,[¶] Shuai Cheng Li,[†] Yehia Mechref,[¶] Haixu Tang,[‡] and Xiaowen Liu[*,§,‖]

[†]*Department of Computer Science, City University of Hong Kong*

[‡]*School of Informatics and Computing, Indiana University Bloomington*

[¶]*Department of Chemistry and Biochemistry, Texas Tech University*

[§]*Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis*

[‖]*Center for Computational Biology and Bioinformatics, Indiana University School of Medicine*

E-mail: xwliu@iupui.edu

Phone: +1-317-278-7613. Fax: +1-317-278-9201

Abstract

Glycosylation is one of the most common post-translational modifications in proteins, existing in about 50% of mammalian proteins. Several research groups have demonstrated that mass spectrometry is an efficient technique for glycopeptide identification. However, this problem is still challenging because of the enormous diversity of glycan structures and the microheterogeneity of glycans. In addition, a glycopeptide may contain multiple glycosylation sites, making the problem complex. Current software tools often fail to identify glycopeptides with multiple glycosylation sites, hence

This is the author's manuscript of the article published in final edited form as:
Zhang, Y., Yu, C.-Y., Song, E., Li, S. C., Mechref, Y., Tang, H., & Liu, X. (2015). Identification of Glycopeptides with Multiple Hydroxylysine O-Glycosylation Sites by Tandem Mass Spectrometry. Journal of Proteome Research, 14(12), 5099–5108. http://doi.org/10.1021/acs.jproteome.5b00299

we present GlycoMID, a graph-based spectral alignment algorithm that can identify glycopeptides with multiple hydroxylysine O-glycosylation sites by tandem mass spectra. GlycoMID was tested on mass spectrometry data sets of the bovine collagen α-(II) chain protein, and experimental results showed that it identified more glycopeptide-spectrum-matches than other existing tools, including many glycopeptides with two glycosylation sites.

Introduction

Glycosylation, existing in about 50% mammalian proteins,[1] plays vital roles in many cellular events, such as signal transduction and receptor activation, and is related to many diseases, including the congenital disorders of glycosylation and cancer.[2,3] There has been increasing demand for techniques and bioinformatics tools for the identification and characterization of glycoproteins.

Glycans, which are attached to proteins (or lipids) in glycosylation, have many different compositions and structures, making the patterns of glycosylation extremely diverse.[2] Even for one specific glycosylation site, there are many glycans that can be attached to the site, which is called microheterogeneity.[4] As a result, it is a challenging problem to identify and characterize glycoproteins.

Mass spectrometry (MS) has been widely used for the identification of protein glycosylation because it is high throughput and capable of sequencing glycopeptides as well as decoding glycan structures.[5–7] In bottom-up MS-based glycosylation analysis, glycoproteins are first digested into short glycopeptides by a protease, and then the resulting glycopeptides are enriched and analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS).[8,9] Fragmentation methods in MS/MS determine the patterns of fragment ions observed in MS/MS spectra of glycopeptides. Collision-induced dissociation (CID) MS/MS spectra contain many glycosidic ions, which involve the breakage of glycan bonds; electron transfer dissociation (ETD) MS/MS spectra contain many ions with peptide backbone frag-

mentation and intact glycosylation structures on the glycosylation sites; higher-energy collisional dissociation (HCD) MS/MS undergoes a similar process as CID, but results in a ladder of mono-, di-, or tri-saccharide oxonium ions and a dominated Y1 ion (peptide+GlcNAc) in spectra.[10–12]

The two main types of glycosylation are N-linked and O-linked glycosylation; other types of glycosylation include C-mannosylation, phosphoglycosylation, and glypiation.[13] N-linked glycosylation involves the attachment of glycans to asparagine or arginine residues, and O-linked glycosylation to serine, threonine, tyrosine, hydroxyproline, or hydroxylysine (HyK) residues.[14] Moreover, N-glycans share a common core, GlcNAc2Man3, and follow biosynthetic pathways in extension;[15] O-glycan attachments have various types, such as glucose (Glc), galactose (Gal), mannose, and fucose.[13]

Several studies have demonstrated that glycan structures can be successfully identified by LC-MS/MS.[16] There are three interconnected problems in analyzing a glycopeptide: determining the glycan structures, sequencing the peptide, and localizing the glycosylation sites. Many software tools have been developed for the first problem, including SimGlycan,[17] GlycoWorkbench,[18] OSCAR,[19] GLYCH,[20] GlycosidIQ[21] and STAT.[22] Most of the tools were designed for the determination of the structures of N-glycans by using CID/HCD MS/MS spectra, which contain glycan fragment ions. There are also many software tools for the second and third problems, such as GlycoPep Grader[23] and GlycoPep Detector,[24] which utilize ETD MS/MS spectra that contain many peptide fragment ions. In addition, some software tools provide a solution to the three interconnected problems, including Peptoonist,[25] GlycoFragwork,[26] GlycoMaster DB,[27] Sweet-Heart,[28] MAGIC,[29] and the one proposed by Cheng et al.[30] In these tools, the first problem is solved by a *de novo* or glycan database search method, and the second and third problems by a peptide database search method. However, the tools can analyze glycopeptides with only one glycosylation site due to the complexity of glycan structures.

In this study, we focus on a special O-linked glycosylation in which Gal or glucose-

galactose (Glc-Gal) moieties are attached to hydroxylysine residues in collagen. Although multiple glycosylation sites may exist in a single peptide,[31,32] most existing software tools fail to identify this type of glycopeptides. To solve this problem, we propose GlycoMID, a spectral alignment algorithm for the identification of glycopeptides with multiple hydroxylysine O-glycosylation sites using MS/MS. In glycopeptide analysis, the determination of glycan structures and the other two tasks: peptide sequencing and glycosylation site localization, can be performed either separately or simultaneously. The tasks are performed separately in MAGIC[29] and simultaneously in GlycoMaster DB,[27] BiOnic,[33] and other peptide identification tools in which glycosylations are treated as variable PTMs. In GlycoMID, the three tasks are also performed simultaneously: an MS/MS spectrum is searched against all possible glycoforms of the peptides in a database to find a glycoform that best explains the spectrum. A limitation of GlycoMID is that it identifies glycopeptides that involve only one type of monosaccharides or several types of monosaccharides with the same mass, such as Glc and Gal. GlycoMID was tested on MS/MS data sets of the collagen α-1(II) chain protein from bovine cartilage, and experimental results showed that it identified more glycopeptides than Mascot[34] and X!Tandem.[35]

Methods

Mass spectrometry experiments

The protein collagen α-1 (II) (CO2A1) chain was acquired from bovine cartilage.[36,37] Pepsinized immunization-grade bovine CO2A1 samples were prepared in either an ammonium bicarbonate buffer (ABC) or phosphate-buffered saline (PBS). The samples prepared in PBS were digested by GluC; the samples prepared in ABC were digested by either trypsin or GluC.

The digested peptides were subjected to LC-MS/MS analysis using a nano-LC system interfaced to an LTQ Orbitrap Velos^{TM} mass spectrometer with a nano-ESI source. Two configurations of the mass spectrometer were used to acquire MS/MS data. For the first

configuration, each full MS scan (resolution 15 000) was followed by CID and HCD MS/MS pair scans analyzing the 8 most intense ions in the MS scan, and the isolation width for collecting parent ions was set as 3 m/z. The first configuration was applied to analyze one sample prepared in PBS and digested by GluC as well as two samples prepared in ABC: one digested by trypsin and the other by GluC. For the second configuration, each MS scan was followed by triplicate scans of CID, HCD, and ETD analysing the 5 highest peaks in the MS scan, and the isolation width for collecting parent ions was set as 4 m/z. The second configuration was applied to analyze two samples prepared in ABC: one digested by trypsin and the other by GluC. For both the configurations, CID and ETD spectra were collected in low resolution in the linear ion trap, and HCD spectra were collected in high resolution in the Orbitrap analyzer.[38] A total of five MS/MS data sets were collected, containing 12 908 CID, 13 986 HCD, and 1 937 ETD MS/MS spectra.[7]

Collagen glycopeptide characterization problem

Unlike common glycosylation sites, O-linked glycosylation on collagen often involves only one or two monosaccharides (Gal or Glc-Gal). In addition, a peptide of collagen may contain multiple glycosylation sites. Because the glycans are short, an MS/MS spectrum of a collagen glycopeptide often contains both common fragment ions, such as b- and y-ions in CID/HCD spectra or c- and $z^{\bullet}$-ions in ETD spectra, and ions generated from a peptide bond cleavage and a glycosidic bond cleavage (Fig. 1).[16,31,39] These ions provide enough information for simultaneous identification of glycosylation sites and glycan structures.

When an MS/MS spectrum is given and the unmodified form of the target peptide is known, the objective of glycosylation analysis is to find a glycoform (glycopeptide) of the peptide that best explains the spectrum. Because of the microheterogeneity of glycopeptides, there are many candidate glycoforms of the peptide whose molecular masses match the precursor mass of the spectrum. Since the number of candidate glycopeptides may be large, how to efficiently find the best glycoform is a challenging problem.

In the proposed algorithm, a peak counting score is used to evaluate glycopeptide-spectrum-matches. The peak counting score between an experimental spectrum and a glycopeptide is the number of matched peak pairs (within an error tolerance) between the spectrum and the theoretical spectrum of the glycopeptide. We formulate the glycopeptide characterization problem on collagen as follows.

Collagen glycopeptide characterization problem Given an MS/MS spectrum, an unmodified peptide, and a number m, find a glycoform of the peptide with m monosaccharides that maximizes the peak counting score between the spectrum and the glycopeptide.

Representing the problem as a graph We will solve the glycopeptide characterization problem using a graph-based method. While both charge 1 and charge 2 ions were used for glycopeptide identification in the experiments, below only charge 1 ions are included to simplify the description of the proposed algorithm. In general, an MS/MS spectrum of a collagen glycopeptide is more complex than that of the unmodified form of the peptide because glycosidic bonds in glycans may be cleaved, increasing the complexity of the spectrum. For example, while the unmodified peptide AGFKEGQKGE has only one form of b_6 ions with charge 1, a glycoform of the peptide with m monosaccharides in a linear structure may have $m+1$ forms of b_6 ions with charge 1 because the fragment ion may contain $0, 1, \ldots,$ or m monosaccharides (Fig. 1). We use only the number of monosaccharides, not the structure, to distinguish fragment ions, that is, two fragment ions with the same charge state, amino acids, and number of monosaccharides, but different glycan structures, are treated as the same.

The proposed method can be applied to CID, HCD and ETD MS/MS spectra. Here HCD MS/MS spectra are used to explain the method. To simplify the analysis, only b- and y-ions are included in the generation of theoretical spectra of glycopeptides. We further assume that monosaccharides in glycans have the same mass.

Let $P = a_1 a_2 \ldots a_n$ be the unmodified form of a peptide and S an MS/MS spectrum of

a glycoform of P with m monosaccharides. Let $b_{i,j}$ and $y_{i,j}$ be the m/z values of the singly charged b_i and y_i ions with j monosaccharides of the glycoform, respectively. We generate two $(n-1)\times(m+1)$ matching matrices B and Y by mapping $b_{i,j}$ and $y_{i,j}$, for $1 \leq i \leq n-1$ and $0 \leq j \leq m$, to the spectrum S (Fig. 2(b)). If $b_{i,j}$ ($y_{i,j}$) matches the m/z value of a peak in S, then $B[i,j] = 1$ ($Y[i,j] = 1$); otherwise, $B[i,j] = 0$ ($Y[i,j] = 0$).

A directed graph containing $n+1$ layers is generated based on the matching matrices with three steps (Fig. 2(c)). First, we generate $n-1$ layers, each containing $m+1$ nodes, in the middle of the graph. The jth node in the ith layer represents the prefix $a_1 a_2 \ldots a_i$ of P with j monosaccharides, denoted by $v_{i,j}$. When the prefix contains j monosaccharides, the suffix $a_{i+1} \ldots a_n$ contains $m-j$ monosaccharides. The peaks matching $b_{i,j}$ and $y_{n-i,m-j}$ are called the *matching peaks* of the node $v_{i,j}$. In addition, the glycan bonds in the prefix or suffix may also be fragmented, resulting in peaks matching $b_{i,0}, \ldots, b_{i,j-1}, y_{n-i,0}, \ldots, y_{n-i,m-j-1}$. These peaks are called the *supporting peaks* of the node. We designed two functions for assigning a weight to a node in the $n-1$ layers: the matching weight function and the combined weight function. The first involves only matching peaks: the weight for the node $v_{i,j}$ is $w_{i,j} = B[i,j] + Y[n-i, m-j]$; the second considers both matching and supporting peaks: the weight $w_{i,j} = B[i,j] + Y[n-i, m-j] + \alpha(\sum_{j'=0}^{j-1} B[i,j'] + \sum_{j'=0}^{m-j-1} Y[n-i,j'])$, where α is a user-specified parameter for supporting peaks. Second, we add layer 0 and layer n. Layer 0 contains only one node $v_{0,0}$, and layer n contains only one node $v_{n,m}$. The weights of the two nodes are 0. Third, directed edges are added between two nodes in neighboring layers. If a glycan can be attached to the ith residue in the peptide, node $v_{i-1,j'}$ is connected to node $v_{i,j}$ by a directed edge if and only if $j - \beta \leq j' \leq j$, where β is the number of monosaccharides in the largest glycan in a collagen peptide. Otherwise, node $v_{i-1,j'}$ is connected to $v_{i,j}$ by a directed edge if and only if $j' = j$.

Each path from $v_{0,0}$ to $v_{n,m}$ corresponds to a glycoform of the peptide. For example, the heaviest path in Fig. 2(c) corresponds to the glycopeptide AGF**k**(Gal)EGQ**k**(Glc-Gal)GE that best explains the peaks in the spectrum (Fig. 3(a)). Compared with the best glycopep-

tide, another candidate glycopeptide AGF**k**(Glc-Gal)EGQ**k**(Gal)GE corresponds to a path with less weight (Fig. 2(c)) and explains less peaks in the spectrum (Fig. 3(b)). When the theoretical spectrum of the glycoform contains only common b and y-ions with charge 1, the matching weight of the path equals the peak counting score between the glycoform and the spectrum. When the theoretical spectrum contains both common b and y-ions with charge 1 and those with glycosidic bond cleavages, the combined weight ($\alpha = 1$) equals the peak counting score. As a result, the glycopeptide characterization problem is reduced to the heaviest path problem in the graph, which can efficiently solved by a dynamic programming algorithm. The number of operations of the algorithm is proportional to nm^2, where n is the length of the peptide and m is the number of monosaccharides in the glycoform.

Database search In peptide identification, peptides are filtered by the precursor mass of the spectrum to speed up database search. Similarly, we use the precursor mass of the spectrum to filter peptides in glycopeptide identification. If the difference between the precursor mass of the spectrum and the molecular mass of a peptide matches the mass of a combination of allowed glycans (within an error tolerance), the peptide is chosen as a candidate. Each candidate peptide is aligned with the spectrum to find the best glycoform for it, and finally the best scoring glycoform among all candidate peptides is reported.

Results

We implemented GlycoMID in C++ and tested it on a desktop with an Intel Core i5 2.3 GHz dual-core CPU and 8 GB memory.

Glycopeptide identification

Proteome Discoverer version 1.4 (Thermo Scientific, San Jose, CA) was used to convert Thermo raw files to MGF files. Following the preprocessing steps in Ref. 7, noise peaks were

removed using a signal-to-noise ratio 1.5, and MS/MS spectra without peaks were removed. After preprocessing, 12 908 CID, 13 186 HCD and 1 937 ETD MS/MS spectra were kept. In addition, the m/z values of the peaks were calibrated based on the differences between their experimental and theoretical m/z values. (See the supplementary material for details.)

In database searches, three types of post-translational modifications (PTMs) were set as variable PTMs: carbamidomethyl on cysteine, hydroxylysine, and hydroxyproline.[40–42] The error tolerance for precursor masses was set as 0.02 Da. Since the CID and ETD spectra were collected in the linear ion trap of the mass spectrometer, the error tolerance for CID and ETD fragment masses was set as 0.5 Da. The error tolerance for HCD fragment masses was also set as 0.5 Da so that the same parameter setting was used for the three types of spectra. The protein database contained the peptides of CO2A1 protein and shuffled decoy peptides of the same size. Two approaches were used to estimate false discovery rates (FDRs) of identified peptide-spectrum-matches. In the first approach, peptide-spectrum-matches with or without glycosylation sites are combined to estimate FDRs. Using this approach GlycoMID identified 4 081 spectra, including 1 901 CID, 1 809 HCD, and 371 ETD spectra, from 28 031 spectra with a 1% spectrum-level FDR. A total of 2 873 unmodified peptide-spectrum-matches and 1 208 glycopeptide-spectrum-matches were identified, including 1 152 matches with one glycosylation site and 56 matches with two glycosylation sites. Of the matches with one glycosylation site, 605 contain one Gal and 547 contain one Glc-Gal. Of the matches with two glycosylation sites, 27 contain two Gals or two Glc-Gals, and 29 contain one Gal and one Glc-Gal. A total of 125 glycopeptides were identified, and the numbers of identified glycopeptides using HCD, CID, and ETD spectrum were 98, 84, and 70, respectively (Fig. 4(a)). In the second approach, we divided the identified peptide-spectrum-matches into three groups with 0, 1, or 2 glycosylation sites, and estimated FDRs of identified peptide-spectrum-matches in the three groups separately. Using the second approach, GlycoMID identified 4 069 peptide-spectrum-matches (CID: 1 899, HCD: 1 825, and ETD: 345), including 2 872 matches without glycosylation, 1 122 matches with 1 glycosylation site, and 75

matches with 2 glycosylation sites, with a 1% spectrum-level FDR.

Coverage of the CO2A1 protein

The CO2A1 protein contains 68 lysine residues, of which 24 residues were identified as glycosylation sites by GlycoMID. By manual investigation, the residue K634 was removed from the list due to the low quality match, and the residue K781 was removed due to the ambiguity of the glycosylation site localization. Of the remaining 22 residues, the residue K929 was glycosylated with Glc-Gal only, and the other 21 were glycosylated with either Gal or Glc-Gal. (See the supplementary material.) All the 22 residues are at the hydroxylysine positions of the Gly-Xaa-HyK motif,[36,37,43] and the glycosylation forms of the sites are consistent with a previous study.[7] The CO2A1 protein possesses 24 possible glycosylation sites with the Gly-Xaa-Hyk motif, including the 22 identified glycosylation sites. Of the remaining two sites, the site K773 was covered by identified peptides without glycosylation sites, and the site K1130 was not covered by any identified peptides. The MS/MS spectra of the peptides covering the site K1130 may lack enough fragment ion peaks, making them unidentifiable.

Because there are various glycosylated forms of the CO2A1 protein, the identified peptides may have different PTMs (including different glycans) on a glycosylation site. Two approaches were utilized to estimate the frequency that the CO2A1 protein has a PTM on a glycosylation site. The first approach is based on spectral counting. Let n_1, n_2, n_3, n_4 be the numbers of identified spectra that support the unmodified lysine residue, the hydroxylysine residue, the hydroxylysine residue with a Gal, and the hydroxylysine residue with a Glc-Gal on a glycosylation site, respectively. The frequency of the unmodified lysine residue on the site is calculated as $\frac{n_1}{n_1+n_2+n_3+n_4}$, and those of other forms of modified residues are computed similarly. In the second approach, the intensity of the precursor ion of an MS/MS spectrum is treated as its weight, and the sums of the weights of the spectra that support the four forms of lysine residues are computed for the estimation.

The frequencies of the four forms of the residues on the 22 identified glycosylation sites

and the site K773 are summarized in Table 1. (See the supplementary material for details of the identified peptides.) K470 and K956 were mainly unmodified or hydroxylated; K287 was only hydroxylated or glycosylated; K299, K308, K374 and K803 were mainly hydroxylated or glycosylated; K731 was mainly glycosylated; K929 was only glycosylated with Glc-Gal. After manual investigation, we found that most peptides covering K929 contain an oxidized methionine residue. Since this PTM was not included in the analysis, we only identified one form of the residue (the hydroxylysine residue with a Glc-Gal) on the site K929.

Comparison of the weight functions

We described two functions in Section "Methods" for assigning weights to nodes in the graph generated from the collagen glycopeptide characterization problem: the matching weight function and the combined weight function. We tested the two functions on all of the five data sets. Using a gradient method, we found that the combined weight function achieved the best performance when $\alpha = 0.65$. With a 1% spectrum-level FDR, the combined weight function approach ($\alpha = 0.65$) identified 4 081 peptide-spectrum-matches, including 1 208 matches with glycosylation sites (CID: 545, HCD: 524, and ETD: 139); the matching weight function approach identified 3 980 peptide-spectrum-matches, including 1 129 matches with glycosylation sites (CID: 489, HCD: 494, and ETD: 136). The reason for the better performance of the combined weight function is that the fragmentation process sometimes breaks both peptide and glycan bonds, resulting in both matching and supporting peaks. Excluding supporting peaks in the matching weight function will deteriorate its performance.

Fragment ions

Because Gal and Glc-Gal are short glycans, it is common to observe fragment ions with a loss of Gal or Glc-Gal. A fragment ion with only one Gal glycosylation site either keeps or loses the Gal. These two forms are called a Gal-ion and a Gal-loss ion, respectively. We computed the frequencies that Gal and Gal-loss ions were observed for b and y-ions with charge 1 or 2

in CID and HCD spectra and for c and $z^{\bullet}$ ions with charge 1 or 2 in ETD spectra, based on the 605 identified peptide-spectrum-matches with one Gal glycosylation site from 271 CID, 275 HCD, and 59 ETD spectra. (See Section "Glycopeptide identification.")

Below Gal b-ions with charge 1 in CID spectra are used to show how to compute the frequency. We generate all theoretical Gal b-ions with charge 1 from the identified glycopeptide-spectrum-matches from CID spectra, and find the theoretical ions that are matched to an experimental peak in the corresponding spectrum. The frequency is computed as the ratio between the number of the matched theoretical ions and the total number of the theoretical ions.

Fig. 5 shows the frequencies of Gal and Gal-loss ions in CID, HCD, and ETD spectra. While the ratio between the frequencies of Gal and Gal-loss ions in ETD spectra is about 7, those in CID and HCD spectra are about 1, supporting previous studies.[10–12]

Similarly, using spectrum-peptide matches with one Glc-Gal glycosylation site, the frequencies of fragment ions with a loss of Glc or Glc-Gal were obtained (Fig. 6). Compared with ETD spectra, HCD and CID spectra contain more fragment ions with a loss of Gal or Glc-Gal, which is consistent with previous findings:[44–49] CID tends to produce fragments of glycopeptides with minimal backbone fragmentation, and ETD tends to break glycopeptides along the peptide backbones, preserving glycans.

Due to the presence of prevalent glycan loss peaks in CID/HCD MS/MS spectra, GlycoMID might report glycopeptides with incorrect glycan compositions and glycosylation sites. To evaluate the performance of GlycoMID on the localization of glycosylation sites for CID/HCD MS/MS spectra with glycan loss peaks, we compared CID/ETD and HCD/ETD spectral pairs in the two data sets using HCD/CID/ETD alternate fragmentation. We analyzed the 36 CID/ETD and 41 HCD/ETD spectral pairs satisfying the condition that both the spectra were mapped to glycoforms of the same peptide that contains multiple lysine residues (candidate glycosylation sites), which may result in errors in the localization of glycosylation sites. Of these spectral pairs, 9 CID/ETD and 12 HCD/ETD spectral pairs

reported different localization results. Manual inspection showed that most localization results reported by the 9 CID and 12 HCD spectra were not reliable because of the absence of key fragment ions with intact modifications.

Combination of CID, HCD and ETD MS/MS spectra

In the MS experiments, two data sets were generated using HCD/CID/ETD alternate fragmentation. The CID, HCD, and ETD spectra in these data sets were combined to improve glycopeptide identification. Using the combined spectra and the parameter settings described in Section "Glycopeptide identification," we identified 1 470 peptide-spectrum-matches, including 543 matches with glycosylation and 927 matches without glycosylation. Of the 1 470 matches, 171 matches with glycosylation and 163 matches without glycosylation were missed by peptide identification using single spectra (Fig. 7). However, the combined approach also missed 111 peptide-spectrum-matches identified by single spectra. By manual inspection, we found that most of the 111 unidentified spectral triplets have only one high quality spectrum in the triplet.

Comparison with Mascot and X!Tandem

We compared GlycoMID and Mascot on the CO2A1 data sets. The input of Mascot was the same MGF files generated for GlycoMID. For Mascot, the error tolerances for precursor and fragment masses were set as 0.02 Da and 0.5 Da, respectively. Fixed and variable PTMs as well as the fragment ions for scoring were set as the same to the analysis in Section "Glycopeptide identification." All the MS/MS spectra were searched against the target-decoy concatenated peptide database used in Section "Glycopeptide identification." With a 1% spectrum level FDR, Mascot identified 3 531 spectra, of which 845 were matched to glycopeptides, A total of 91 glycopeptides were identified by Mascot, including 69, 63 and 38 glycopeptides from HCD, CID, and ETD MS/MS spectra, respectively (Fig. 4(b)). All the glycopeptides identified by Mascot were also identified by GlycoMID, and GlycoMID

identified more peptide-spectrum-matches than Mascot (Fig. 8). While GlycoMID includes various fragment ions with and without glycan loss into the combined weight function, Mascot does not use all the fragment ions in the computation of similarity scores. That might be the main reason for the better performance of GlycoMID.

Although the scoring function of GlycoMID includes supporting peaks, it does not utilize the information of peak intensities. After manually inspection, we found out that CID and HCD MS/MS spectra contain more low-intensity noise peaks than ETD MS/MS spectra in the test data sets. These noise peaks were treated the same as high intensity signal peaks in the scoring function of GlycoMID. As a result, GlycoMID reported more decoy random glycopeptide-spectrum-matches for CID/HCD spectra than ETD ones. Because of these decoy matches, more CID/HCD target identifications of GlycoMID were filtered out than ETD target identifications with the same FDR. This might be the reason that the improvement of GlycoMID for CID/HCD spectra is less significant than that for ETD spectra. To further improve the performance of GlycoMID, a better scoring function should be designed to include both supporting peaks and patterns of the intensities of various fragment ion peaks. In addition, the performance of GlycoMID can be improved by using high accuracy and high resolution MS/MS spectra with small errors in the m/z values of fragment ion peaks.

The data sets were also analyzed using X!Tandem, which reported less identifications than Mascot and GlycoMID. The data sets contain peptides with hydroxylation of lysine and two types of glycosylation of hydroxylysine residues, X!Tandem cannot handle three types of variable PTMs on the same amino acid residue, resulting in the miss of many identifications.

Discussion and conclusions

We proposed an algorithm GlycoMID for the identification of glycopeptides with O-linked glycosylation sites. The experiments on the CO2A1 data sets demonstrated that GlycoMID

efficiently identified glycopeptides in the protein. Based the identified peptide-spectrum-matches, a total of 22 lysine residue of CO2A1 were hydroxylated and glycosylated with Glc-Gal or Gal moieties, which is consistent with previous studies. We also showed that combining CID, HCD, and ETD spectra improved on the number of identified peptide-spectrum-matched compared with single spectra.

A main limitation of the implementation of GlycoMID is that it identified glycopeptides with only one type of monosaccharides or several types of monosaccharides with the same mass. A glycopeptide may contain two or more types of monosaccharides with different masses. When a length n glycopeptide contains k monosaccharides with t different masses, the number of possible b_i ions of the peptide is an exponential function proportional to k^t, and the number of nodes in the graph representing the peptide and its possible glycosylation sites is proportional to nk^t, making the algorithm inefficient. To further speed up the algorithm, one method is to reduce the search space by allowing only a list of commonly observed glycans, each corresponding to a combination of monosaccharides. When the number of allowed glycans is g and the number of glycosylation sites is s, the number of nodes in the graph representing the peptide and its possible glycosylation sites is proportional ng^s, which is generally smaller than nk^t. The number of operations of the algorithm is proportional to ng^s.

GlycoMID was designed for the identification of glycopeptides, not for the computation of localization scores of glycosylation sites. For an MS/MS spectrum, it reports only a best scoring glycopeptide with its glycosylation sites and structures, not localization scores of the identified glycosylation sites. Many methods have been proposed for the localization of PTMs, such as A-score,[50] PTM score,[51] PhosphoRS[52] and Mascot Delta Score.[53] These methods compute a PTM localization score based on the similarity scores between an MS/MS spectrum and a peptide with different PTM sites. Similarly, the proposed algorithm can be modified to report the similarity scores between an MS/MS spectrum and a glycopeptide with different glycosylation sites, which can be utilized to compute localization scores of

glycosylation sites.

Because of the complexity of glycosylation and the similarity of glycoforms from the same peptide, multiple glycoforms with the same molecular mass may not be separately by LC, resulting in multiplexed MS/MS spectra of a mixture of the glycoforms. For such a spectrum, GlycoMID reports a glycoform with the highest similarity score, not multiple glycoforms. Computational methods have been proposed for the identification of multiple proteoforms from a multiplexed spectrum in histone protein studies.[54,55] These methods can be applied to identify multiple glycoforms from a multiplexed MS/MS spectrum.

Supporting Information

The CO2A1 data sets are publicly available at MassIVE (http://massive.ucsd.edu/) with accession id MSV000079174. The software GlycoMID is available at http://proteomics.informatics.iupui.edu/software/glycomid/.

Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Steen, P. V. d.; Rudd, P. M.; Dwek, R. A.; Opdenakker, G. Concepts and principles of O-linked glycosylation. *Critical reviews in biochemistry and molecular biology* **1998**, *33*, 151–208.

(2) Ohtsubo, K.; Marth, J. D. Glycosylation in cellular mechanisms of health and disease. *Cell* **2006**, *126*, 855–867.

(3) Dallas, D. C.; Martin, W. F.; Hua, S.; German, J. B. Automated glycopeptide analysis-review of current state and future directions. *Briefings in bioinformatics* **2013**, *14*, 361–374.

(4) Harmon, B. J.; Gu, X.; Wang, D. I. Rapid monitoring of site-specific glycosylation microheterogeneity of recombinant human interferon-γ. *Analytical chemistry* **1996**, *68*, 1465–1473.

(5) Harazono, A.; Kawasaki, N.; Itoh, S.; Hashii, N.; Ishii-Watabe, A.; Kawanishi, T.; Hayakawa, T. Site-specific N-glycosylation analysis of human plasma ceruloplasmin using liquid chromatography with electrospray ionization tandem mass spectrometry. *Analytical biochemistry* **2006**, *348*, 259–268.

(6) Borges, C. R.; Jarvis, J. W.; Oran, P. E.; Nelson, R. W. Population studies of vitamin D binding protein microheterogeneity by mass spectrometry lead to characterization of its genotype-dependent O-glycosylation patterns. *Journal of proteome research* **2008**, *7*, 4143–4153.

(7) Song, E.; Mechref, Y. LC–MS/MS identification of the O-glycosylation and hydroxylation of amino acid residues of collagen α-1 (II) chain from bovine cartilage. *Journal of proteome research* **2013**, *12*, 3599–3609.

(8) Nilsson, J.; Rüetschi, U.; Halim, A.; Hesse, C.; Carlsohn, E.; Brinkmalm, G.; Larson, G. Enrichment of glycopeptides for glycan structure and attachment site identification. *Nature methods* **2009**, *6*, 809–811.

(9) Ongay, S.; Boichenko, A.; Govorukhina, N.; Bischoff, R. Glycopeptide enrichment and separation for protein glycosylation analysis. *Journal of separation science* **2012**, *35*, 2341–2372.

(10) Wuhrer, M.; Catalina, M. I.; Deelder, A. M.; Hokke, C. H. Glycoproteomics based on tandem mass spectrometry of glycopeptides. *Journal of Chromatography B* **2007**, *849*, 115–128.

(11) Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E.; Shabanowitz, J.;

Hunt, D. F. The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2006**, *1764*, 1811–1822.

(12) Segu, Z. M.; Mechref, Y. Characterizing protein glycosylation sites through higher-energy C-trap dissociation. *Rapid Communications in Mass Spectrometry* **2010**, *24*, 1217–1225.

(13) Spiro, R. G. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* **2002**, *12*, 43R–56R.

(14) Yan, A.; Lennarz, W. J. Unraveling the mechanism of protein N-glycosylation. *Journal of Biological Chemistry* **2005**, *280*, 3121–3124.

(15) Lowe, J. B.; Marth, J. D. A genetic approach to mammalian glycan function. *Annual review of biochemistry* **2003**, *72*, 643–691.

(16) Harvey, D. J. *Proteomic analysis of glycosylation: structural determination of N-and O-linked glycans by mass spectrometry*; Future Drugs Ltd London, UK, 2005.

(17) Albanese, J.; Glueckmann, M.; Lenz, C. SimGlycan^{TM} Software*: a new predictive carbohydrate analysis tool for MS/MS data. *Appl Biosystems* **2010**,

(18) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *Journal of proteome research* **2008**, *7*, 1650–1659.

(19) Lapadula, A. J.; Hatcher, P. J.; Hanneman, A. J.; Ashline, D. J.; Zhang, H.; Reinhold, V. N. Congruent strategies for carbohydrate sequencing. 3. OSCAR: An algorithm for assigning oligosaccharide topology from MSn data. *Analytical chemistry* **2005**, *77*, 6271–6279.

(20) Tang, H.; Mechref, Y.; Novotny, M. V. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* **2005**, *21*, i431–i439.

(21) Joshi, H. J.; Harrison, M. J.; Schulz, B. L.; Cooper, C. A.; Packer, N. H.; Karlsson, N. G. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* **2004**, *4*, 1650–1664.

(22) Gaucher, S. P.; Morrow, J.; Leary, J. A. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Analytical chemistry* **2000**, *72*, 2331–2336.

(23) Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. GlycoPep Grader: A web-based utility for assigning the composition of N-linked glycopeptides. *Analytical chemistry* **2012**, *84*, 4821–4829.

(24) Zhu, Z.; Hua, D.; Clark, D. F.; Go, E. P.; Desaire, H. GlycoPep detector: a tool for assigning mass spectrometry data of N-linked glycopeptides on the basis of their electron transfer dissociation spectra. *Analytical chemistry* **2013**, *85*, 5023–5032.

(25) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H. R.; Dell, A. Automated N-glycopeptide identification using a combination of single- and tandem-MS. *Journal of proteome research* **2007**, *6*, 3995–4005.

(26) Mayampurath, A.; Yu, C.-Y.; Song, E.; Balan, J.; Mechref, Y.; Tang, H. Computational framework for identification of intact glycopeptides in complex samples. *Analytical chemistry* **2013**, *86*, 453–463.

(27) He, L.; Xin, L.; Shan, B.; Lajoie, G. A.; Ma, B. GlycoMaster DB: software to assist the automated identification of N-linked glycopeptides by tandem mass spectrometry. *Journal of proteome research* **2014**, *13*, 3881–3895.

(28) Wu, S.-W.; Liang, S.-Y.; Pu, T.-H.; Chang, F.-Y.; Khoo, K.-H. Sweet-Heart-an integrated suite of enabling computational tools for automated MS2/MS3 sequencing and identification of glycopeptides. *Journal of proteomics* **2013**, *84*, 1–16.

(29) Lynn, K.-S.; Chen, C.-C.; Lih, T.-S. M.; Cheng, C.-W.; Su, W.-C.; Chang, C.-H.; Cheng, C.-Y.; Hsu, W.-L.; Chen, Y.-J.; Sung, T.-Y. MAGIC: an automated N-linked glycoprotein identification tool using a Y1-ion pattern matching algorithm and in silico MS2 approach. *Analytical Chemistry* **2015**, *87*, 2466–2473.

(30) Cheng, K.; Chen, R.; Seebun, D.; Ye, M.; Figeys, D.; Zou, H. Large-scale characterization of intact N-glycopeptides using an automated glycoproteomic method. *Journal of proteomics* **2014**, *110*, 145–154.

(31) Peter-Katalinić, J. Methods in enzymology: O-Glycosylation of proteins. *Methods in enzymology* **2005**, *405*, 139–171.

(32) Wiederschain, G. Y. Essentials of glycobiology. *Biochemistry (Moscow)* **2009**, *74*, 1056–1056.

(33) Bern, M.; Cai, Y.; Goldberg, D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal Chem* **2007**, *79*, 1393–400.

(34) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(35) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–7.

(36) Miller, E. J.; Lunde, L. G. Isolation and characterization of the cyanogen bromide peptides from the αl (II) chain of bovine and human cartilage collagen. *Biochemistry* **1973**, *12*, 3153–3159.

(37) Chung, E.; Miller, E. J. Collagen polymorphism: characterization of molecules with the chain composition [α1 (III)] 3 in human tissues. *Science* **1974**, *183*, 1200–1201.

(38) Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Muller, M.; Viner, R.; Schwartz, J.; Remes, P.; Belford, M.; Dunyach, J. J.; Cox, J.; Horning, S.; Mann, M.; Makarov, A. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol Cell Proteomics* **2012**, *11*, O111 013698.

(39) Quan, L.; Liu, M. CID, ETD and HCD Fragmentation to Study Protein Post-Translational Modifications. *Modern Chemistry & Applications* **2013**,

(40) Gelse, K.; Pöschl, E.; Aigner, T. Collagens-structure, function, and biosynthesis. *Advanced drug delivery reviews* **2003**, *55*, 1531–1546.

(41) Shoulders, M. D.; Raines, R. T. Collagen structure and stability. *Annual review of biochemistry* **2008**, *78*, 929–958.

(42) Yamauchi, M.; Sricholpech, M. Lysine post-translational modifications of collagen. *Essays in biochemistry* **2012**, *52*, 113–133.

(43) Butler, W. T.; Miller, E. J.; Finch Jr, J. E.; Inagami, T. Homologous regions of collagen α1 (I) and α1 (II) chains: Apparent clustering of variable and invariant amino acid residues. *Biochemical and biophysical research communications* **1974**, *57*, 190–195.

(44) Mirgorodskaya, E.; Hassan, H.; Clausen, H.; Roepstorff, P. Mass spectrometric determination of O-glycosylation sites using β-elimination and partial acid hydrolysis. *Analytical chemistry* **2001**, *73*, 1263–1269.

(45) Balog, C. I.; Mayboroda, O. A.; Wuhrer, M.; Hokke, C. H.; Deelder, A. M.; Hensbergen, P. J. Mass spectrometric identification of aberrantly glycosylated human apolipoprotein C-III peptides in urine from Schistosoma mansoni-infected individuals. *Molecular & Cellular Proteomics* **2010**, *9*, 667–681.

(46) Kelleher, N. L.; Zubarev, R. A.; Bush, K.; Furie, B.; Furie, B. C.; McLafferty, F. W.; Walsh, C. T. Localization of labile posttranslational modifications by electron capture dissociation: the case of γ-carboxyglutamic acid. *Analytical chemistry* **1999**, *71*, 4250–4253.

(47) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 9528–9533.

(48) Mormann, M.; Paulsen, H.; Peter-Katalinić, J. Electron capture dissociation of O-glycosylated peptides: radical site-induced fragmentation of glycosidic bonds. *European Journal of Mass Spectrometry* **2005**, *11*, 497–511.

(49) Sihlbom, C.; van Dijk Härd, I.; Lidell, M. E.; Noll, T.; Hansson, G. C.; Bäckström, M. Localization of O-glycans in MUC1 glycoproteins using electron-capture dissociation fragmentation mass spectrometry. *Glycobiology* **2009**, *19*, 375–381.

(50) Beausoleil, S. A.; Villén, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology* **2006**, *24*, 1285–1292.

(51) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006**, *127*, 635–648.

(52) Taus, T.; Köcher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and confident phosphorylation site localization using phosphoRS. *Journal of Proteome Research* **2011**, *10*, 5354–5362.

(53) Savitski, M. M.; Lemeer, S.; Boesche, M.; Lang, M.; Mathieson, T.; Bantscheff, M.;

Kuster, B. Confident phosphorylation site localization using the Mascot Delta Score. *Molecular & Cellular Proteomics* **2011**, *10*, M110.003830.

(54) Baliban, R. C.; DiMaggio, P. A.; Plazas-Mayorca, M. D.; Young, N. L.; Garcia, B. A.; Floudas, C. A. A novel approach for untargeted post-translational modification identification using integer linear optimization and tandem mass spectrometry. *Molecular & Cellular Proteomics* **2010**, *9*, 764–79.

(55) DiMaggio, J., P. A.; Young, N. L.; Baliban, R. C.; Garcia, B. A.; Floudas, C. A. A mixed integer linear optimization framework for the identification and quantification of targeted post-translational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. *Molecular & Cellular Proteomics* **2009**, *8*, 2527–43.

Table

Table 1: The frequencies of the unmodified lysine residue, the hydroxylysine residue, the hydroxylysine residue with a Gal, and the hydroxylysine residue with a Glc-Gal on the 22 identified glycosylation sites and the site K773 in the CO2A1 protein. Two frequencies are reported for each form of the lysine residue at each glycosylation site. The first is estimated by the spectral counting method, and the second by the method that incorporates the intensities of precursor ions.

Glycosylation	Relative frequency (%)			
site	Unmodified K	HyK	Gal-HyK	Glc-Gal-HyK
K287	0.00/0.00	65.94/76.72	15.68/11.44	18.38/11.84
K299	2.79/0.83	54.98/46.73	27.09/36.25	15.14/16.19
K308	4.19/1.13	71.85/78.55	11.98/11.83	11.98/8.49
K374	0.32/0.25	29.62/19.37	40.44/45.90	29.62/34.48
K419	9.52/11.39	49.21/44.06	15.87/13.80	25.40/30.75
K452	22.64/55.65	24.53/27.29	22.64/9.53	30.19/7.53
K464	10.00/8.24	54.44/66.97	16.67/13.72	18.89/11.07
K470	50.00/35.20	33.55/50.22	9.87/10.47	6.58/4.11
K527	33.97/53.47	19.62/9.43	22.97/21.04	23.44/16.06
K542	34.38/26.47	0.00/0.00	28.12/27.28	37.50/46.25
K608	50.00/55.66	7.14/26.83	21.43/11.22	21.43/6.29
K620	46.18/31.55	37.82/31.66	7.27/6.68	8.73/30.14
K731	9.56/1.91	3.48/1.18	20.87/14.57	66.09/82.34
K764	36.54/27.83	25.00/44.19	17.31/12.35	21.15/15.63
K773	87.06/86.65	12.94/13.35	0.00/0.00	0.00/0.00
K803	6.32/9.83	46.84/40.24	19.47/10.20	27.37/39.73
K848	19.32/12.21	35.88/48.46	32.27/32.88	12.53/6.45
K857	15.00/12.97	70.00/71.44	7.50/9.59	7.50/6.00
K884	14.29/5.42	4.76/2.41	7.14/14.72	73.81/77.45
K929	0.00/0.00	0.00/0.00	0.00/0.00	100.00/100.00
K956	48.76/64.36	40.50/32.91	8.26/2.33	2.48/0.40
K1055	38.32/33.63	34.58/56.88	16.82/6.40	10.28/3.09
K1118	29.55/18.63	61.82/80.42	4.54/0.49	4.09/0.46

Figures

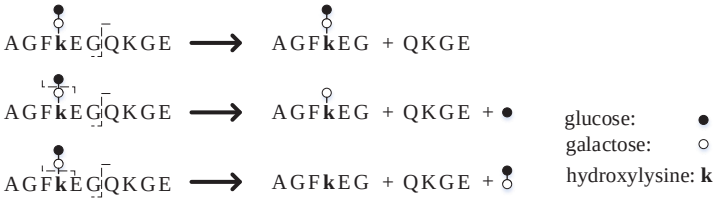
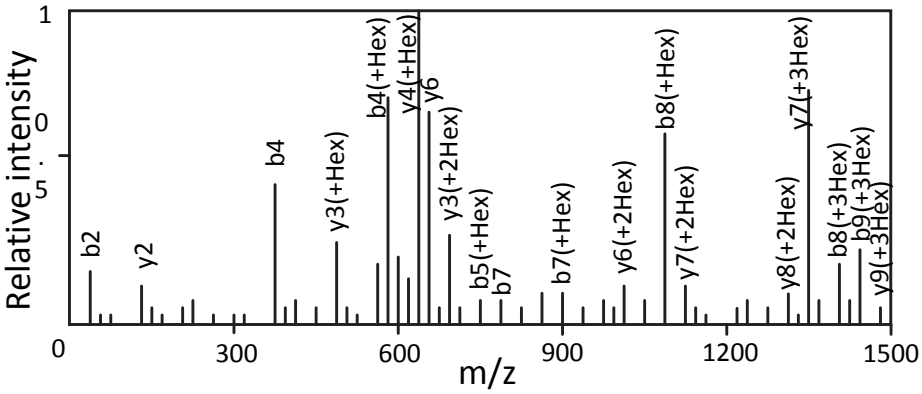


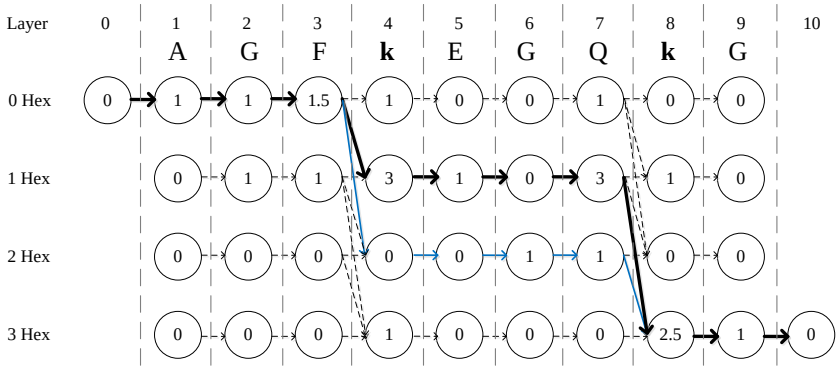
Figure 1: Possible fragment ions generated from a glycopeptide AGF**k**(Glc-Gal)EGQKGE, in which **k** is a hydroxylysine residue.



(a)

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad Y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(b)



(c)

Figure 2: Conversion from the glycopeptide characterization problem to the heaviest path problem in a graph. (a) An HCD MS/MS spectrum annotated with possible fragment ions of glycoforms of AGF**k**EGQ**k**GE with 3 Gal/Glc monosaccharides, where **k** represents a hydroxylysine residue. (b) The matching matrices B and Y generated from the annotated spectrum. (c) A graph is generated based on the matrices B and Y using the combined weight function ($\alpha = 0.5$). The heaviest path, shown in bold, corresponds to the glycoform AGF**k**(Gal)EGQ**k**(Glc-Gal)GE. The path from layer 0 to layer 10 containing the blue subpath corresponds to the glycoform AGF**k**(Glc-Gal)EGQ**k**(Gal)GE.

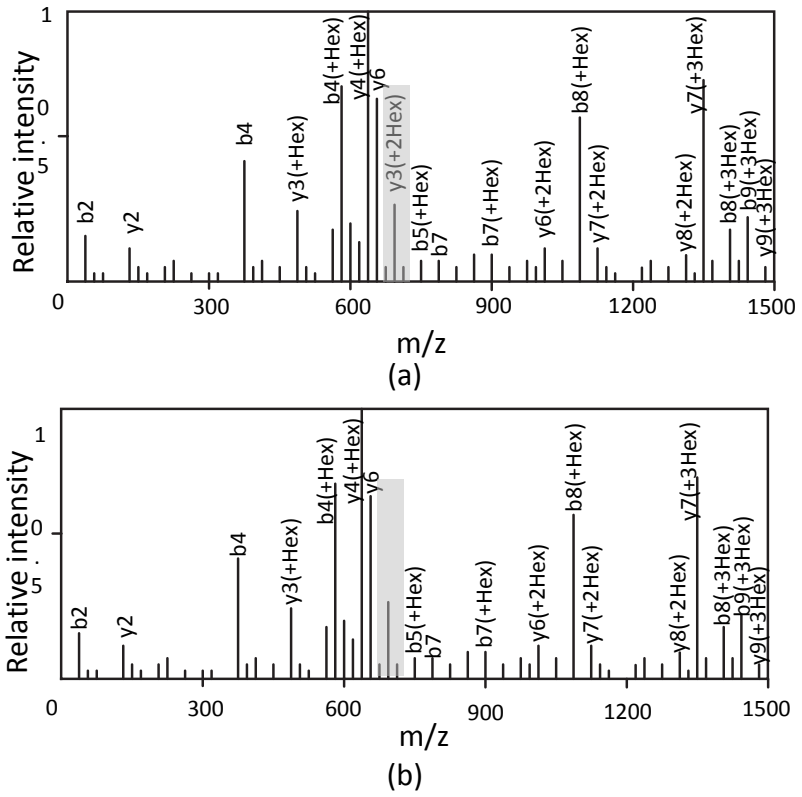
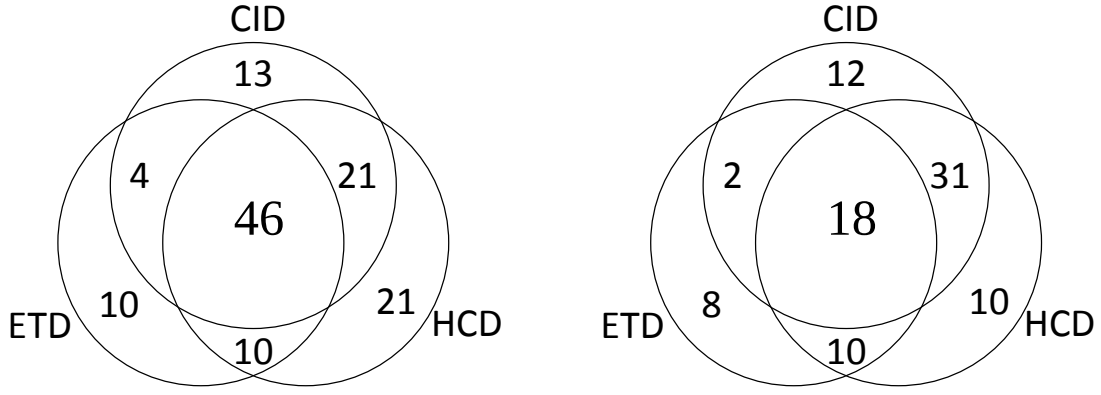


Figure 3: Comparison between two glycopeptides matched to an MS/MS spectrum. (a) The HCD MS/MS spectrum in Fig. 2(a) is annotated with fragment ions of glycopeptide AGF**k**(Gal)EGQ**k**(Glc-Gal)GE, which corresponds to the heaviest path in Fig. 2(c). (b) The same spectrum is annotated with fragment ions of glycopeptide AGF**k**(Glc-Gal)EGQ**k**(Gal)GE, which corresponds to the path with the blue subpath in Fig. 2(c) with less weight. The peak in the shade area is annotated by AGF**k**(Gal)EGQ**k**(Glc-Gal)GE, not by AGF**k**(Glc-Gal)EGQ**k**(Gal)GE.



(a) GlycoMID (b) Mascot

Figure 4: The numbers of glycopeptides identified with different fragmentation methods: (a) GlycoMID; (b) Mascot.

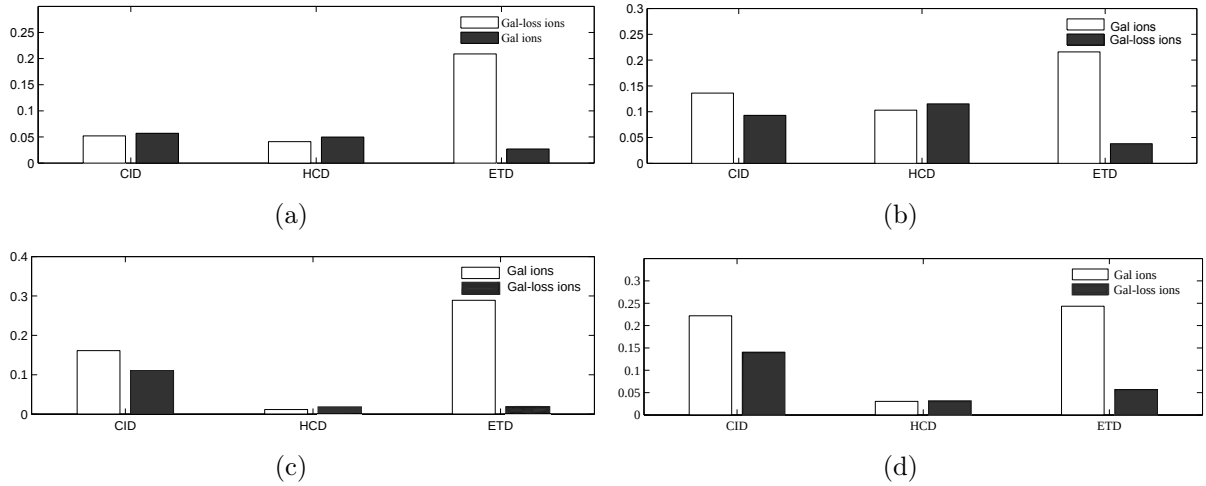


Figure 5: Frequencies of Gal and Gal-loss fragment ions in the identified 605 glycopeptide-spectrum-matches with only one Gal glycosylation sites, including 271 CID, 275 HCD, and 59 ETD spectra. (a) Prefix fragment ions with charge 1 (b-ions for CID and HCD spectra and c-ions for ETD spectra). (b) Suffix fragment ions with charge 1 (y-ions for CID and HCD spectra and $z^{\bullet}$-ions for ETD spectra). (c) Prefix fragment ions with charge 2 (b-ions for CID and HCD spectra and c-ions for ETD spectra). (d) Suffix fragment ions with charge 2 (y-ions for CID and HCD spectra and $z^{\bullet}$-ions for ETD spectra).

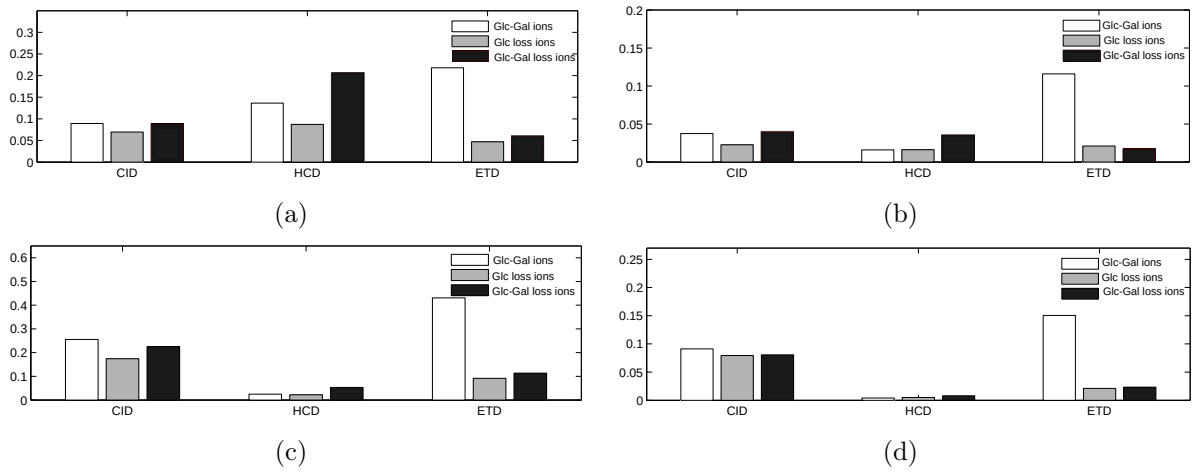


Figure 6: Frequencies of various fragment ions in the identified 547 glycopeptide-spectrum-matches with only one Glc-Gal glycosylation sites, including 251 CID, 228 HCD, and 68 ETD spectra. Three types of fragment ions that contain only one glycosylation site with an Glc-Gal are compared: fragment ions keeping the Glc-Gal (Glc-Gal ions), losing the Glc (Glc-loss ions), and losing the Glc-Gal (Glc-Gal-loss ions). (a) Prefix fragment ions with charge 1 (b-ions for CID and HCD spectra and c-ions for ETD spectra). (b) Suffix fragment ions with charge 1 (y-ions for CID and HCD spectra and $z^{\bullet}$-ions for ETD spectra. (c) Prefix fragment ions with charge 2 (b-ions for CID and HCD spectra and c-ions for ETD spectra). (d) Suffix fragment ions with charge 2 (y-ions for CID and HCD spectra and $z^{\bullet}$-ions for ETD spectra).

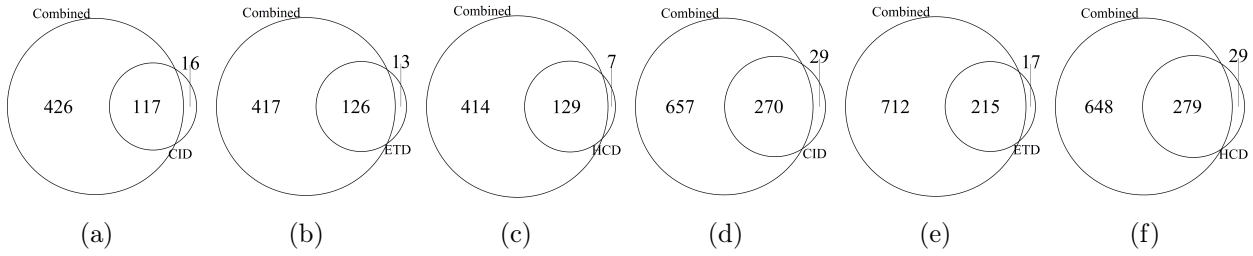


Figure 7: Comparison of the numbers of peptide-spectrum-matches identified by single spectra and by combining multiple spectra on the data sets with CID/HCD/ETD triplets: (a-c) glycopeptide-spectrum-matches; (d-f) peptide-spectrum-matches without glycosylation.

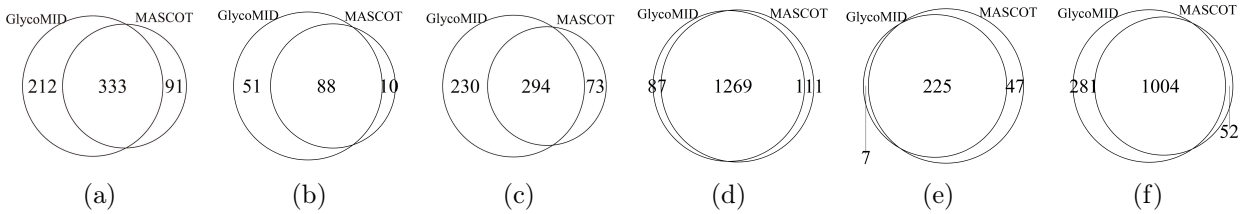


Figure 8: Comparison between the numbers of peptide-spectrum-matches reported by Mascot and GlycoMID with a 1% spectrum level FDR. (a-c) Glycopeptide-spectrum-matches identified from CID, ETD, and HCD spectra, respectively. (d-f) Peptide-spectrum-matches without glycosylation identified from CID, ETD, and HCD spectra, respectively.