Journal of proteome research

pubs.acs.org/jpr

Article

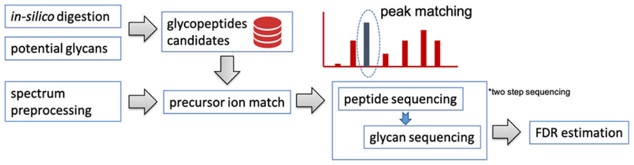
GlycoHybridSeq: Automated Identification of N-Linked Glycopeptides Using Electron Transfer/High-Energy Collision Dissociation (EThcD)

Rui Zhang, Jianhui Zhu, David M. Lubman, Yehia Mechref, and Haixu Tang*

Cite This: *J. Proteome Res.* 2021, 20, 3345−3352

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Glycosylation is one of the most common post-translational modifications (PTM) occurring in a large variety of proteins with important biological functions in human and other higher organisms. Liquid chromatography tandem mass spectrometry (LC-MS/MS) has been routinely used to characterize site-specific protein glycosylation at high throughput in complex glycoproteomic samples. Recently, electron transfer/high-energy collision dissociation (EThcD) was introduced for glycopeptide identification, which offers rich structural information on glycopepides with the fragment ions from the cleavages of both the glycan and the peptide backbone. Herein, we present the software GlycoHybridSeq for automated interpretation of EThcD-MS/MS spectra from glycoproteomic data using a customized scoring function, which enables the functionalities of identifying glycopeptides, characterizing glycosylation sites, and distinguishing some isomeric glycans. We evaluate GlycoHybridSeq on glycoproteomic data collected for cancer biomarker discovery. The results showed that it achieved comparable or better performance than that of Byonic and MSFragger. GlycoHybridSeq is released as an open source software and is ready to be used in large-scale glycoproteomic data analyses.

KEYWORDS: *glycoproteomics, tandem mass spectrometry, EThcD, algorithm, software tool, glycopeptide identification, GlycoHybridSeq*

INTRODUCTION

Glycosylation is a post-translational modification (PTM) occurring in a large variety of proteins involved in important biological functions, such as immune response, host−pathogen interactions, cellular differentiation and adhesion, and signal transductions, in higher animals like humans.[1] The aberrant alteration of glycan structure is implicit with the malfunction of cells and possesses potential significance for the early medical diagnosis of complex human diseases including cancer.[2−4] Liquid chromatography tandem mass spectrometry (LC-MS/MS) has been commonly applied to the analyses of glycomic and glycoproteomic samples, aiming to identify glycopeptide biomarkers from human bodily fluids (e.g., blood samples) that are associated with cancers, in particular the cancer of different organs or different types.[5−8]

Various fragmentation modes have been used for glycopeptide identification using tandem mass spectrometry (MS/MS). Collision-induced dissociation (CID) leads to the cleavages of glycosidic bonds in the glycans without breaking the peptide backbone, resulting in the series of Y-ions that provides the structural information on the glycans in glycopeptides. On the other hand, electron transfer dissociation (ETD) allows the fragmentation of the peptide backbones of glycopeptides while retaining the intact glycan, and thus enables the sequencing of the peptide. Finally, higher energy collision induced dissociation (HCD) generates oxonium ions with high mass accuracy from glycopeptides in addition to the Y-ions, which are indicative the monosaccharide composition of the glycans. Because these methods provide complementary information about glycopeptides,[9] they are often combined for glycopeptide identification, in particular for the characterization of the site-specific protein glycosylations in complex glycoproteomic samples (e.g., from human blood samples).[10−12] However, the employment of multiple fragmentation modes requires multiple scans of the same ions (sometimes even multiple analyses of the same sample due to the instrument constraints), and the sensitivity of the analyses may be sacrificed as a trade-off.

Recently, electron transfer/high-energy collision dissociation (EThcD) was offered as an alternative option to generate rich structural information on glycopepides, featuring the fragment ions from the cleavages of both the glycan and the peptide backbone.[13,14] Because it requires only a single scan for a putative glycopeptide ion, EThcD became increasingly popular in glycoproteomics, especially for biomarker discovery.[15,16] However, unlike for the other conventional fragmentation methods, the software support for glycopeptide identification

Received: March 26, 2021
Published: May 19, 2021

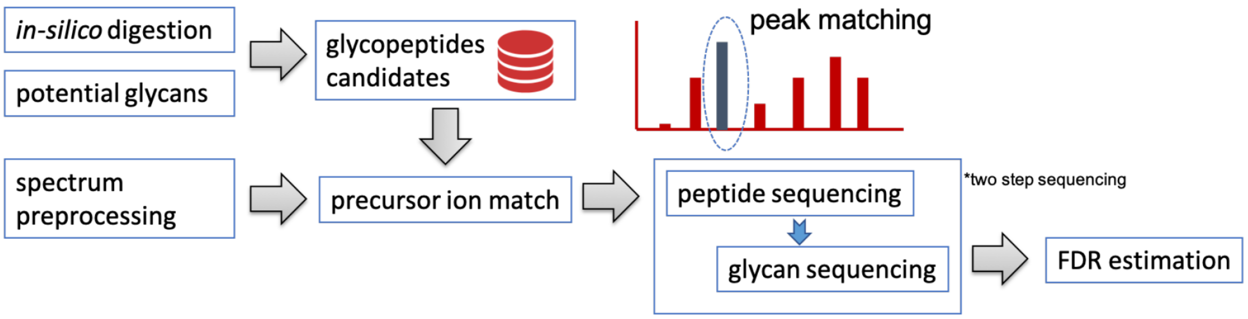


Figure 1. Workflow of the GlycoHybridSeq algorithm.

from EThcD-MS/MS spectra is still limited. To the best of our knowledge, the commercial software Byonic[17] provides an option specifically designed for EThcD spectra, while MSFragger-glyco allows for glycopeptide identification using EThcD[18] even though its scoring scheme was not optimized for EThcD spectra.

In this paper, we present the software tool GlycoHybridSeq, which extended the glycopeptide identification algorithm previously implemented in GlycoSeq,[19] for analyzing glycoproteomic data acquired using EThcD. GlycoHybridSeq incorporates a scoring function designed for EThcD-MS/MS spectra, and enables the identification of glycopeptides, the characterization of the glycosylation sites, and the ability to distinguish many isomeric glycans all automatically. We evaluate GlycoHybridSeq on glycoproteomic data collected for cancer biomarker discovery. The results showed it achieved comparable or better performance than Byonic and MSFragger. GlycoHybridSeq is released as an open source software and is ready to be used in large-scale glycoproteomic data analyses.

■ METHODS

Using electron transfer/high-energy collision dissociation (EThcD), glycopeptides may be cleaved within their peptide backbones, generating c/z and b/y fragment ions containing intact glycans, or at the glycosidic bonds in the glycans, generating b/y fragment ions containing the intact peptide backbone. Therefore, we devised a scoring scheme for assessing if a putative glycopeptide-spectrum matches (GSMs; with the matched precursor mass) that takes into consideration these fragment ions.

In this paper, we focused on the identification of N-linked glycopeptides from their EThcD-MS/MS spectra. To speed up the process of scoring and ranking GSMs, we precompute the fragment ions for all potential N-glycans with up to a certain maximum number of monosaccharide residues (by default, #HexNAc $\leq$ 12, #Hex $\leq$ 12, #Fuc $\leq$ 5, and #NeuAc $\leq$ 4) according to the rule of biosynthetic pathways. We note that in GlycoHybridSeq, we did not consider the unlikely N-glycans, e.g., those missing part of pentamer core. During the precomputation, the N-glycans are represented as an array, in which each bit represents the number of a specific monosaccharide in the core or each branch. For example, a quaternary complex-type N-glycan is represented as a 20 dimensional array (e.g., [2, 3, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0]), the first three dimensions representing two N-acetylglucosamines (GlcNAc), three mannoses (Man), and one fucose (Fuc) in the pentamer core, respectively, the next dimension representing zero bisected GlcNAc, the next 12 dimensions representing one GlcNAc, zero galactose (Gal) and zero sialic acid (e.g., NeuAc) in each of the four branches of the N-glycan. Using this representation of N-glycans, GlycoHybridSeq employs a dynamic programming algorithm to find the N-glycopeptide (i.e., an N-glycan attached to a peptide backbone derived from the in silico digestion of a target protein) that receives the maximum score of matched fragment ions with a given input spectrum.

In contrast to O-glycans, the human N-glycans have a well-defined structure, which are formed with GlcNAc3Man2 core and (one of) three different types of branches (i.e., complex, hybrid, and high mannose). On the basis of this kind of defined structure, the software GlycoHybridSeq automatically generates the complete list of N-glycans following rules of glycosidic linkage as defined by glycan synthetic pathways[20] to eliminate improbable glycan structures and build reasonable glycan trees. An example of the growth of glycans can be found in the Supporting Information.

The workflow of GlycoHybridSeq is shown as Figure 1. Each experimental EThcD-MS/MS spectrum is first searched against all putative N-glycans and peptides through the precursor ion match to identify candidate glycopeptide-spectrum matches (GSMs). Each candidate GSM is then subject to the algorithms for peptide sequence match and glycan sequencing to obtain the peptide and glycan matching score, respectively. The GSMs are evaluated by using a scoring scheme that integrates these scores, and the top scored GSMs for each EThcD-MS/MS spectrum is reported along with the matching scores. Finally, a procedure to estimate the false discovery rate (FDR) is implemented GlycoHybridSeq, which will determine a score threshold corresponding to a desirable FDR cutoff (e.g., 0.01). In the next sections, we describe the details of these components implemented in GlycoHybridSeq.

Precursor Ion Match

The precursor ion match is the first step to identify putative GSMs in GlycoHybridSeq, which matches the mass of a precursor ion with that of a putative glycopeptide. GlycoHybridSeq enumerates all putative N-glycans, and for each of them, searches for the candidate peptide backbones with the corresponding mass (i.e., within the mass tolerance from the precursor ion mass subtracting the glycan mass). The candidate peptides are obtained from the in silico digestion (e.g., by trypsin) of the proteins of interest (e.g., the human haptoglobin in the study used here for evaluation; see below). We used the *bucket search* algorithm for the fast precursor ion match, in which the running time is $O(N_G)$, where N_G is the total number of putative N-glycans (by default GlycoHybridSeq considers five types of N-glycans with up to 12 HexNAc, 12 Hex, 5 Fuc, and 4 NeuAc monosaccharides). Each experimental EThcD-MS/MS spectrum will form candidate GSMs with multiple glycopeptides that match its precursor mass, which will be used for peptide and glycan sequencing as well as the GSM scoring in the subsequent steps.

Bucket Search

To search a particular mass within a given mass tolerance from the mass values in a target set, we employed a bucket search algorithm. The list of buckets is first created, in which each value in the target set is put into one of the buckets based on an indexing function. For a given mass tolerance in Dalton, the indexing function is defined by

$$i = \left\lfloor \frac{\mathrm{val} - \mathrm{low}}{\mathrm{tol}} \right\rfloor \tag{1}$$

where i is the index of the bucket, tol is the mass tolerance, val is a mass value of interest, and low is the smallest value in the target set. Here, the size of the bucket is proportional to the mass tolerance so that any two mass values within the tolerance is assigned into the same bucket or the adjacent ones. As a result, to find the mass values within a tolerance from an input mass value, it is sufficient to look up mass values in the same bucket and its two adjacent buckets.

Similarly, for the tolerance given in the units of parts per million (PPM), the indexing function is defined by

$$i = \lfloor \log_{1/1-d/10^6}^{\mathrm{val/low}} \rfloor \tag{2}$$

Note that the boundaries of the buckets define by the indexing function form a geometric sequence: $a_i = \frac{1}{1-d/10^6}^{i} a_0$, where a_0 = low is the minimum mass value in the target set. Therefore, for any value of interest a^*, if $a_{i-1} < a^* < a_i$, the difference of $|a^* - a_0|/a_0$ is always smaller than d ppm, indicating the mass tolerance is smaller than the bucket size.

Peptide Sequence Match

GlycoHybridSearch considers the c/z and b/y fragment ions resulting from peptide backbone fragmentation[13] for matching peptide sequences. For each candidate of GSMs obtained in the glycan search (including the N-glycan and the corresponding peptide backbone), the algorithm first computes the theoretical mass values of all putative fragment ions, and searches each of them in the list of peaks in the input EThcD-MS/MS spectrum using the bucket search algorithm. The matched peaks are stored in a table for each GSM, which will be used for glycan sequencing and GSM scoring as described below.

Glycan Sequencing

The glycan sequencing aims to characterize the branching structure of the N-glycans attached to the peptide backbone. GlycoHybridSeq implemented a dynamic programming (DP) algorithm similar to the one used in GlycoSeq[19] but with various modifications to improve its performance. The algorithm considers a group of GSMs corresponding to the same experimental EThcD-MS/MS spectrum, and employs a priority queue to store all theoretical fragment ions resulting from in silico glycosidic cleavages of these glycopeptides. Here, each theoretical fragment ion is associated with its mass value (used as the key in the priority queue) and a list of different *fragmented* glycopeptides (containing the intact peptide backbone and the attached glycan fragment after the cleavage) of the same mass as well as the corresponding ion matching information that is updated during the glycan sequencing algorithm (see below). The theoretical Y1 ions (i.e., the peptide backbones each attached with a single GlcNAc) are first pushed into the priority queue, followed by the fragment ions with greater attached glycan fragments created through the dynamic programming algorithm (see below), using the mass value as the key. The property of priority queue ensures that the algorithm always processes the fragment ions with the smallest mass value at each time.

To match the theoretical fragment ions from the group of GSMs to those in the EThcD-MS/MS spectrum, a fragment ion is retrieved from the top of the priority queue, and searched against the list of experimental peaks using the bucket search based on its mass value and the expected mass tolerance (the resolution of the MS/MS scans). Here, the masses of experimental peaks are obtained after the deconvolution of the MS/MS spectrum using

$$\mathrm{mass_{peak}} = (m/z - \mathrm{mass_{ion}}) \times c \tag{3}$$

where c is a charge (less than or equal to the precursor ion charge), m/z is the observed mass-to-charge-ration of the peak, $\mathrm{mass_{peak}}$ is the mass value of theoretical fragment ion, and $\mathrm{mass_{ion}}$ is the mass of a proton. If the theoretical fragment ion is matched with an experimental peak, the theoretical fragment ion and its matched peak are recorded. The unmatched fragment ions are experimental peaks and are marked as missing. To speed up the algorithm, if five or more fragment ions are marked as missing, the corresponding glycopeptide will be discarded for further processing.

When an experimental peak is matched with a theoretical fragment ion, the ion matching information on the fragment ion (corresponding to a list of fragmented glycopeptides) is updated. The ion matching score of the fragment ion is then computed based on the ion matching information,

$$\mathrm{score} = \sum_i \log(I_i) \tag{4}$$

where I_i is the intensity of experimental peaks in the EThcD-MS/MS spectrum matched with the top scored fragmented glycopeptide among the list of those with the same mass. Notably, in each step of the dynamic programming, only the top scored glycopeptide fragment is retained for each theoretical fragment ion, because the other glycopeptide fragment of the same mass but with a lower score will not lead to a complete glycopeptide with a higher overall score. We note that GlycoHybridSeq considered all three types of N-glycans, including the complex, the high-mannose, and the hybrid types.

After each step of dynamic programming, the glycopeptide fragments grow with an additional monosaccharide following the rule of biosynthetic pathways. The resulting fragment ions corresponding the fragments (along with the ion matching information) are then pushed into the priority queue. Because these new fragment ions always have greater mass values than those already in the queue, they will be processed after the prior ones. The growth of the glycopeptide fragments terminates if they are beyond the user-defined maximum size of the glycopeptides (by default with at most 12 HexNAc, 12 Hex, 5 Fuc, and 4 NeuAc). The dynamic programming algorithm continues until all fragment ions in the priority queue are processed. At the end of the algorithm, for each EThcD-MS/MS spectrum, the top scored glycopeptide corresponding to its precursor ion (i.e., the fragment ion with the maximum mass) that is compatible with the precursor ion match results is retained as a candidate GSM for further analyses.

Assessment of Glycopeptide-Spectrum Matches (GSMs)

To determine the most likely glycopeptide, the potential glycopeptides are scored according to

$$\text{score} = \frac{\sum_i \log(A_i) \sum_j \log(B_j)}{\sum_k \log(I_k)} \quad (5)$$

where A_i is the intensity of the peak i matched in glycan sequencing and B_j is the intensity of the peak j matched in peptide sequence match, while I_k represents the intensity of every peak k in the entire input spectrum.

False Discovery Rate (FDR) Estimation

We implemented a target-decoy search strategy to estimate the false discovery rate (FDR), following the approach reported by Zhu et al.[21] This approach was used in their glycopeptides FDR analysis that accounts for the target-to-decoy ratio that is not 1:1 as in the conventional peptide identification protocol. The decoy peptide database is constructed using the reverse protein sequences, and the FDR is computed based on the GSMs with the target and decoy peptides as backbones, respectively. By assuming the probability of hits for incorrect target assignments are approximately the same to that of decoys, FDR can be computed as[21]

$$\text{FDR} = \frac{N_d}{N_{\text{total}}}\left(1 + \frac{M_t}{M_d}\right) \quad (6)$$

where N_d is the number of identified GSMs with the decoy peptide backbones, while N_{total} is the total number of identified GSMs with the target or decoy peptide backbones, and M_t and M_d are the numbers of target and decoy peptides, respectively.

GlycoConverter

To retrieve accurate precursor ion information (e.g., precursor charge and the isotopic peaks), GlycoConverter is developed to preprocess the MS1 spectra. The Graphic User Interface (GUI) of GlycoConverter is implemented in C# within the .net framework and Windows Presentation Foundation (WPF) architecture. The software takes as input the raw data from Thermo Fisher MS instruments (in .raw format) by using the MSFileReader library, and generate the human readable output file in Mascot generic format (MGF) or mzML format. The mzML is a XML based data format,[22] developed by a joint effort under HUPO-PSI,[23] which enables the sharing of MS data and can be used by other software tools such as MSFragger.[24] In GlycoConverter, the precursor charge is automatically assigned according to the deconvoluted isotopic peaks for multiply charged ions by using the Patterson routine.[25] To retrieve the theoretical isotopic distribution of a given precursor ion, the BRAIN algorithm[26] is employed. The monoisotopic peaks are then estimated by using the *Average* model of the isotopic envelope[27] generated based on the atomic composition of glycopeptides.

Implementation

Given the input MS/MS data (in mzML or .raw format), the protein sequence database (in FASTA format), and a optional list of glycan candidates, GlycoHybridSeq searches for all potential N-glycopeptides for each input spectrum. To obtain all the potential glycopeptides, the protein sequences are in silico digested to peptides by user-defined proteases (e.g., trypsin) with a maximum missing cleavage (2 by default). Only the peptides containing the sequence motif of N-glycosylation site (i.e., AsnXSer or AsnXThr, where X is any amino acid except proline) are considered for glycopeptide identification. The carbamidomethylation of cysteine is assumed for all peptides.

For each input spectrum, GlycoHybridSeq reported the top scored N-glycopeptides with the score higher than a threshold associated with a user-defined FDR cutoff (0.01 by default). The presence of terminal fucoses and sialic acids are derived if the corresponding peaks specific to the glycans are observed. The user may manually define the searching parameters. The output of the software includes the peptide sequence and glycan structure of the identified glycopeptides as well as the associated scores.

Evaluation

We evaluated GlycoHybridSeq using the LC-MS/MS data set acquired from glycoproteomic samples using EThcD (PRIDE ID: PXD011239) in comparison with Byonic[28] and MSFragger.[24] Briefly, in this study, the haptoglobin (Hp) was immunopurified from serum samples of 5 patients with early stage hepatocellular carcinoma (HCC), 5 patient liver cirrhosis, and 5 healthy subjects, which subsequently was trypsin/GluC digested. Glycopeptides were enriched using HILIC TopTips and analyzed by using LC-EThcD. Trypsin and GluC were selected as the proteases with a maximum of two missed cleavages. The whole data set contains 30 raw files (2 experimental replicates for each human subject), and a total of 450 149 EThcD-MS/MS spectra.

The data analyses using GlycoHybridSeq and other software (Byonic and MSFragger) was performed with precursor ion mass tolerance of 10 ppm and the fragment ion mass tolerance of 0.01 Da. The results were filtered at 1% FDR. To further examine the performance of GlycoHybridSeq, they are also compared with results of two most widely used sequencing software MSFragger[24] and Byonic,[28] respectively.

The Byonic is used in the same manner as described in our previous work.[15] Briefly, all spectra were analyzed by using Byonic (Protein Metrics, San Carlos, CA) incorporated in the Proteome Discoverer 2.1 (Thermo). The default Byonic glycan database consists of 164 mammalian N-glycans, and an additional set of 15 N-glycans reported in the literature for human haptoglobin were also used. Trypsin and GluC were selected as enzymes with a maximum of two missed cleavages allowed. Results were filtered at 1% FDR and a confidence threshold of Byonic score >100. It is worth noting that, by default, Byonic filters its protein list at 1% FDR or after 20 decoy proteins, whichever comes last.[28] Also, Byonic incorporated a special algorithm for the target-decoy strategy to estimate and control the FDR at the peptide-spectrum match (PSM) and the protein levels simultaneously[29]

The MSFragger (v3.1.1) is used with the default parameters for glycopeptide identification (i.e., NGlyco-hybrid.params), provided by the software with a slight modifications to incorporate the digestion of both trypsin and GluC. There is no glycan database explicitly used by MSFragger, while the spectra (mzML format) were analyzed using a total of 182 mass offsets (representing putative protein glycosylations). MSFragger computes the score from the number of matched fragments and their intensity that is used by PeptideProphet for FDR analysis.[18] Philosopher filters results at 1% PSM, both

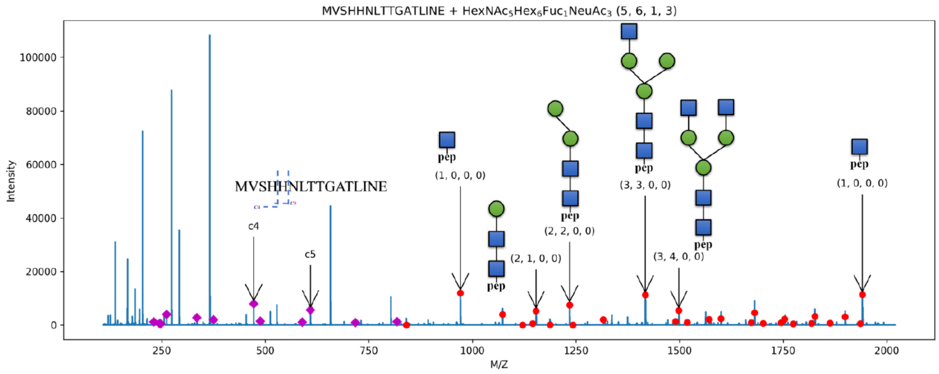


Figure 2. An example of annotated glycopeptide spectrum, where peaks annotated as b/c/y/z ions are colored as purple, and those annotated as Y ions as red.

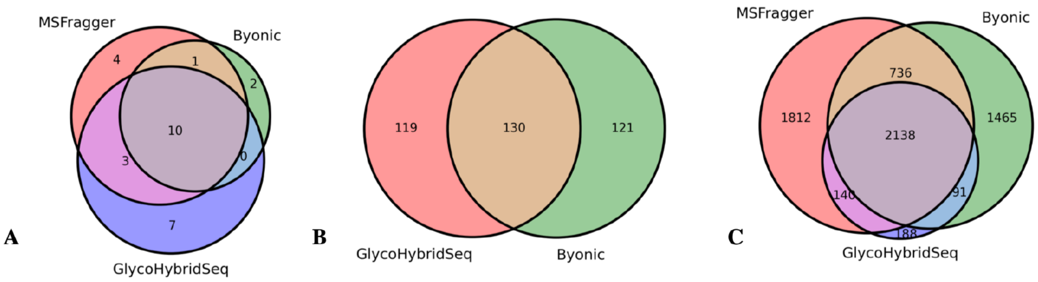


Figure 3. Comparison of (A) unique peptide sequences, (B) unique glycopeptides, and (C) spectra identified by GlycoHybridSeq, Byonic, and MSFragger, respectively.

peptide and protein levels, followed by subsequent filtering of the protein list.[18]

■ RESULTS

GlycoHybridSeq Software

The GlycoHybridSeq is implemented with a graphic user interface (GUI) that takes as input an EThcD-MS/MS data set (in MGF or .raw format) and a database of protein sequences (in Fasta format), and output the identified glycopeptide for each input spectrum (as shown in Figure S1). The software also provides an interface for users to adjust the searching parameters, including mass tolerance, digestion conditions (i.e., enzyme, maximum number of miscleavages, minimal peptide length), other post-translational modifications (PTMs, e.g., oxidation), FDR cutoff, types of N-Glycans, maximum number of monosaccharides, and number of threads (as shown in Figure S1). To improve the performance of searching, it is recommended to consider a relative low desirable value for the maximum number of monosaccharides that is close to actual glycan compositions and a high number of threads when available.

After preprocessing the EThcD-MS/MS spectra and peptide sequences, GlycoHybridSeq filters out the potential glycopeptides by comparing the theoretical and observed precursor ion *m*/*z* within a given mass tolerance (default at 10 ppm). The algorithm then searches each spectrum for any peaks matched to the theoretical fragment ions resulting from the fragmentation of peptide backbone (i.e., the b/c/y/z ions) and glycan (i.e., Y ions), respectively. It scores potential glycopeptides based on the intensities of matched peaks, as described in the Methods section. Figure 2 illustrates the matching process using the EThcD-MS/MS spectrum of the glycopeptide MVSHHNLTTGATLINE attached with the glycan $HexNAc_5$-Hex_6-Fuc_1-$NeuAc_3$ (5, 6, 1, 3). Notably, this spectrum is also matched against another candidate, the glycopeptide NLFLNHSENATAK attached with $HexNAc_9$-Hex_9 (9, 9, 0, 0) that shares a similar precursor mass but a different peptide backbone. However, the fragment ions from the peptide backbone (NLFLNHSENATAK) fragmentation did not match any experimental peak in spectrum, and thus the candidate glycopeptide was not considered in the subsequent scoring.

Comparison with Byonic and MSFragger

To evaluate the glycopeptides identification by GlycoHybridSeq, a glycoproteomic data set of serum haptoglobin for cancer biomarker discovery were used to identify the site-specific intact N-glycopeptides.[15] The data were acquired from the patients of hepatocellular carcinoma (HCC) and liver cirrhosis, and healthy controls, respectively, including a total of 30 raw data files consisting of 450 149 EThcD-MS/MS spectra. The glycopeptides identified from all these MS/MS spectra by GlycoHybridSeq were compared with those by using two other software tools, MSFragger[24] and Byonic.[28] We note that MSFragger only reported the sequences of peptide backbones and the derived mass of the intact glycans, but not the glycan composition information. Hence, only the peptide backbones are compared among three software tools, as shown in Figure 3A. GlycoHybridSearch identified glycopeptides from 20 different peptides, Byonic identified glycopeptides from 13 different peptides, while MSFragger identified glycopeptides from 18 different peptides, among which 10 peptides were identified by all three software tools (see Supplementary Table S1 for details). Interestingly, GlycoHybridSeq and MSFragger both identified glycopeptides from three peptide backbones that were not identified by Byonic, indicating GlycoHybridSeq may identify extra glycopeptides.

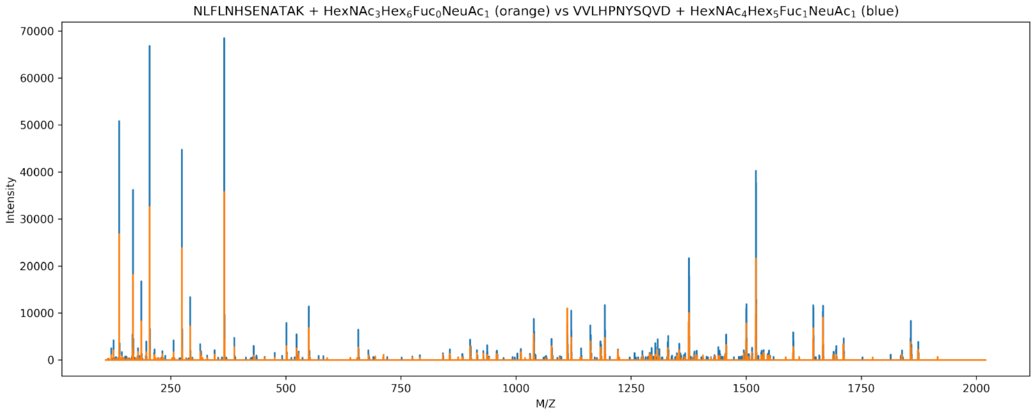


Figure 4. Two EThcD spectra with high cosine similarity were identified as two different glycopeptides by Byonic.

To assess the confidence of spectrum searching of glycopeptides, the MS/MS spectra identified as glycopeptides are also compared among the software tools. As shown in Figure 3C, a total of 2557 spectra were identified by GlycoHybridSeq, 4826 spectra by MSFragger, and 4430 spectra by Byonic. Although MSFragger and Byonic identified more spectra, a large portion of the identifications, 1812 spectra for MSFragger and 1465 spectra for Byonic, respectively, are not identified by the other two searching tools, indicating the additional identified spectra may not be highly confident. On the other hand, most of GlycoHybridSeq identified spectra (92.6%) are identified by at least one other tool (MSFragger or Byonic).

To further analyze the confidence of glycopeptide identifications, we computed the cosine similarities between the experimental spectra that are identified as glycopeptides by each of the three tools, and then clustered using the similarities cutoff of 0.9. We expect the highly similar spectra in the same cluster should be identified as the same glycopeptides. In fact, only 14 out of 224 clusters (6.2%) of spectra identified by GlycoHybridSeq contains those identified as different glycopeptides; in comparison, 16 out of 496 clusters (3.2%) and 66 out of 343 clusters (19.2%) contain the spectra identified as different glycopeptides by MSFragger and Byonic, respectively. Furthermore, among the 14 clusters containing different glycopeptides identified by GlycoHybridSeq, the glycopeptides always share the same peptide backbones, while only the attached N-glycans are isomeric due to distinct numbers of Fucose and NeuAc (because the mass difference between two Fucs and one NeuAc is close to one Dalton). On the other hand, 16 and 66 clusters containing different glycopeptides identified by MSFragger and Byonic, respectively, do not always share the same peptide backbones. For example, Byonic identified two very similar spectra (with the precursor masses of 1111.4684 and 1111.8010, respectively, and the cosine similarity of 0.981) as two different glycopeptides, as shown in Figure 4. Because these two spectra are quite similar, it is unlikely that they result from different glycopeptides. Additional examples of these highly similar spectra and their identifications are presented in the Supporting Information.

■ DISCUSSION

In this paper, we present the open source software GlycoHybridSeq, which is specifically designed for fast and confident glycopeptide identification from LC-EThcD-MS/MS data. Compared to the other fragmentation methods such as CID and ETD, EThcD simultaneously produces unbiased fragment ions resulting from glycan and peptide fragmentation, and thus reveals complete sequence information on glycopeptides.[13] By taking advantage of EThcD, GlycoHybridSeq enables a high throughput identification of N-glycopeptides that contain larger and more complex glycans than O-linked glycopeptides. GlycoHybridSeq achieved performance gaining through a few optimizations based on the specific property of the EThcD fragmentation of glycopeptides. For example, by first sequencing the peptide backbone, many glycopeptide candidates matching with the precursor ion mass but not any peptide backbone fragment ions were eliminated from further consideration; by employing a dynamic programming algorithm, the glycan sequencing can be performed efficiently. Finally, GlycoHybridSeq is implemented in the multithread model that considerably reduce the running time (see Supplementary Figure S2).

GlycoHybridSeq offers an opportunity for highly confident assignment of glycopeptides. The algorithm derives the N-glycan structures based on the rule of biosynthetic pathways, and as a result, it reported only the plausible N-glycan isomers. For example, a total of six glycopeptide isomers correspond to the glycan composition $HexNAc_2$-Hex_1-Fuc_1-$NeuAc_0$ (2, 3, 1, 0), while GlycoHybridSeq reports one plausible N-glycopeptide (i.e., with the glycans containing the pentamer core and one fucose). This approach not only improves the running time, but also avoids false positive N-glycopeptide identifications. Moreover, the scoring function used in GlycoHybridSeq ensures the matching of sufficiently intensive fragment ions resulting from both the peptide and glycans fragmentations, and thus gives higher confidence of the identification results under a low FDR cutoff (i.e., 1%). This is evident from the results that most (92.6%) identified spectra identified by GlycoHybridSeq are also identified by Byonic and/or MSFragger.

Unlikely other tools relying on preprocessed MS/MS data (e.g., by MSFileReader, Thermo Fisher Scientific Inc.) for the precursor information, GlycoHybridSeq implemented its own method for precursor retrieval. In particular, it implemented an independent algorithm (also implemented in GlycoConverter as a standalone tool for converting .raw files into MGF or mzML formats) to compute precursor ion *m*/*z* and charges, which allows it to eliminate some noisy spectra. For example, some MS/MS spectra correspond to no observed peaks in their parent MS spectra, and thus the precursor ion charges arbitrarily assigned to them (e.g., assuming a default charge) are reliable. In fact, a small fraction (17) of glycopeptides were identified by Byonic from such spectra. By employing the Patterson method,[25] it is possible to compute a more accurate charge state (for example, see Figure S3). We note that in GlycoConverter, the monoisotopic peaks were computed by using the average model generated based on the atomic composition of glycopeptides[30] instead of the widely used model generated from peptides composition.[27] In practice, the difference of the two models is small and did not affect the glycopeptide identification (see Figure S4).

GlycoHybridSeq offers a user-friendly GUI. The software itself is complementary to GlycoSeq,[19] which aimed for N-glycopeptide identification from CID/HCD spectra. The source code of GlycoHybridSeq can be found on Github at https://github.com/ruizhang84/GlycoHybridSeq for C++ version on all platforms, at https://github.com/ruizhang84/GlycoSeqApp for Windows version with the GUI support, and at https://github.com/ruizhang84/GlycoConverter for the standalone GlycoConverter tool. It is worth mentioning that the software uses the common target-decoy search approach[31] to estimate the FDR based on the reverse sequence of proteins. To further improve the FDR estimation, it may be beneficial to build decoys that include pesudoglycans that can mimic the false positive of glycans. Sun et al. reported an approach for N-linked glycan identification that constructed a decoy glycan database based on the target glycan structures.[32] Shipman et al. reported a decoy glycopeptide generator (DecoyDeveloper) that can produce a high volume of decoys with low mass differences using a database of 245 biologically relevant glycans.[33] However, these methods for FDR estimation will require further validation and will be pursued in our further endeavors.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00245.

> Figure S1: The snapshot of the GlycoHybridSeq GUI; Figure S2: The running time of GlycoHybridSeq on Linux Manjaro, CPUi5−10500; Figure S3: The charge states computed by different methods; Figure S4: The difference precursor ion *m*/*z* of Avergine model with glycopeptides and peptides compositions; Figure S5: Two EThcD spectra with high cosine similarity were identified as two different glycopeptides by Byonic; Figure S6: Two EThcD spectra with high cosine similarity were identified as two different glycopeptides by MSFragger; Figure S7: The example of growth of N-Glycan with glycosidic linkage as defined by glycan synthetic pathways; Table S1: Table of identified peptides by softwares (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Haixu Tang − *Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana 47405, United States;* Phone: +1 (812) 856-1859; Email: hatang@indiana.edu

Authors

Rui Zhang − *Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0003-0083-1165

Jianhui Zhu − *Department of Surgery, University of Michigan Medical Center, Ann Arbor, Michigan 48109, United States;* orcid.org/0000-0002-0051-7777

David M. Lubman − *Department of Surgery, University of Michigan Medical Center, Ann Arbor, Michigan 48109, United States;* orcid.org/0000-0001-7731-0232

Yehia Mechref − *Department of Chemistry and Biochemistry, Texas Tech University, Lubbock, Texas 79409, United States;* orcid.org/0000-0002-6661-6073

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00245

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the NIH (1R01 GM093322-05, 1R01 GM112490-01, and R01GM49500) and National Cancer Institution (1R01 CA160254 and U01 CA225753).

■ REFERENCES

(1) Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H. *Essentials of Glycobiology*, 3rd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2009; Chapter 9.

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

(3) Pinho, S. S.; Reis, C. A. Glycosylation in cancer: mechanisms and clinical implications. *Nat. Rev. Cancer* **2015**, *15*, 540−555.

(4) Stowell, S. R.; Ju, T.; Cummings, R. D. Protein glycosylation in cancer. *Annu. Rev. Pathol.: Mech. Dis.* **2015**, *10*, 473−510.

(5) Hanash, S. M.; Pitteri, S. J.; Faca, V. M. Mining the plasma proteome for cancer biomarkers. *Nature* **2008**, *452*, 571−579.

(6) Mereiter, S.; Balmaña, M.; Campos, D.; Gomes, J.; Reis, C. A. Glycosylation in the era of cancer-targeted therapy: where are we heading? *Cancer Cell* **2019**, *36*, 6−16.

(7) Wang, M.; Zhu, J.; Lubman, D. M.; Gao, C. Aberrant glycosylation and cancer biomarker discovery: a promising and thorny journey. *Clin. Chem. Lab. Med.* **2019**, *57*, 407−416.

(8) Zhu, J.; Warner, E.; Parikh, N. D.; Lubman, D. M. Glycoproteomic markers of hepatocellular carcinoma-mass spectrometry based approaches. *Mass Spectrom. Rev.* **2019**, *38*, 265−290.

(9) Mechref, Y. Use of CID/ETD mass spectrometry to analyze glycopeptides. *Curr. Protoc. Protein Sci.* **2012**, *68*, 12−11.

(10) Mayampurath, A. M.; Wu, Y.; Segu, Z. M.; Mechref, Y.; Tang, H. Improving confidence in detection and characterization of protein N-glycosylation sites and microheterogeneity. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 2007−2019.

(11) Mayampurath, A.; Yu, C.-Y.; Song, E.; Balan, J.; Mechref, Y.; Tang, H. Computational framework for identification of intact glycopeptides in complex samples. *Anal. Chem.* **2014**, *86*, 453−463.

(12) Mayampurath, A.; Song, E.; Mathur, A.; Yu, C.-y.; Hammoud, Z.; Mechref, Y.; Tang, H. Label-free glycopeptide quantification for biomarker discovery in human sera. *J. Proteome Res.* **2014**, *13*, 4821−4832.

(13) Yu, Q.; Wang, B.; Chen, Z.; Urabe, G.; Glover, M. S.; Shi, X.; Guo, L.-W.; Kent, K. C.; Li, L. Electron-transfer/higher-energy collision dissociation (EThcD)-enabled intact glycopeptide/glycoproteome characterization. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 1751−1764.

(14) Riley, N. M.; Malaker, S. A.; Driessen, M. D.; Bertozzi, C. R. Optimal dissociation methods differ for N-and O-glycopeptides. *J. Proteome Res.* **2020**, *19*, 3286−3301.

(15) Zhu, J.; Chen, Z.; Zhang, J.; An, M.; Wu, J.; Yu, Q.; Skilton, S. J.; Bern, M.; Ilker Sen, K.; Li, L.; et al. Differential quantitative determination of site-specific intact N-glycopeptides in serum haptoglobin between hepatocellular carcinoma and cirrhosis using LC-EThcD-MS/MS. *J. Proteome Res.* **2018**, *18*, 359−371.

(16) Zhu, J.; Huang, J.; Zhang, J.; Chen, Z.; Lin, Y.; Grigorean, G.; Li, L.; Liu, S.; Singal, A. G.; Parikh, N. D.; et al. Glycopeptide Biomarkers in Serum Haptoglobin for Hepatocellular Carcinoma Detection in Patients with Nonalcoholic Steatohepatitis. *J. Proteome Res.* **2020**, *19*, 3452−3466.

(17) Lee, L. Y.; Moh, E. S.; Parker, B. L.; Bern, M.; Packer, N. H.; Thaysen-Andersen, M. Toward automated N-glycopeptide identification in glycoproteomics. *J. Proteome Res.* **2016**, *15*, 3904−3915.

(18) Polasky, D. A.; Yu, F.; Teo, G. C.; Nesvizhskii, A. I. Fast and comprehensive N-and O-glycoproteomics analysis with MSFragger-Glyco. *Nat. Methods* **2020**, *17*, 1125−1132.

(19) Yu, C.-Y.; Mayampurath, A.; Zhu, R.; Zacharias, L.; Song, E.; Wang, L.; Mechref, Y.; Tang, H. Automated glycan sequencing from tandem mass spectra of N-linked glycopeptides. *Anal. Chem.* **2016**, *88*, 5725−5732.

(20) Hamilton, B. S.; Wilson, J. D.; Shumakovich, M. A.; Fisher, A. C.; Brooks, J. C.; Pontes, A.; Naran, R.; Heiss, C.; Gao, C.; Kardish, R.; et al. A library of chemically defined human N-glycans synthesized from microbial oligosaccharide precursors. *Sci. Rep.* **2017**, *7*, 1−12.

(21) Zhu, Z.; Su, X.; Go, E. P.; Desaire, H. New glycoproteomics software, GlycoPep Evaluator, generates decoy glycopeptides de novo and enables accurate false discovery rate analysis for small data sets. *Anal. Chem.* **2014**, *86*, 9212−9219.

(22) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Röompp, A.; Neumann, S.; Pizarro, A. D.; et al. mzML-a community standard for mass spectrometry data. *Mol. Cell. Proteomics* **2011**, *10*, R110.000133.

(23) Orchard, S.; Hermjakob, H. The HUPO proteomics standards initiative-easing communication and minimizing data loss in a changing world. *Briefings Bioinf.* **2007**, *9*, 166−173.

(24) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **2017**, *14*, 513−520.

(25) Senko, M. W.; Beu, S. C.; McLafferty, F. W. Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 52−56.

(26) Dittwald, P.; Claesen, J.; Burzykowski, T.; Valkenborg, D.; Gambin, A. BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Anal. Chem.* **2013**, *85*, 1991−1994.

(27) Senko, M. W.; Beu, S. C.; McLaffertycor, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.

(28) Bern, M.; Kil, Y. J.; Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinf.* **2012**, *40*, 13−20.

(29) Bern, M. W.; Kil, Y. J. Two-dimensional target decoy strategy for shotgun proteomics. *J. Proteome Res.* **2011**, *10*, 5296−5301.

(30) Klein, J. A. *Algorithms for Integrated Analysis of Glycomics and Glycoproteomics by LC-MS/MS*; 2019. https://open.bu.edu/handle/2144/37091.

(31) Elias, J. E.; Gygi, S. P. *Proteome Bioinformatics*; Springer, 2010; pp 55−71.

(32) Sun, W.; Liu, Y.; Zhang, K. An approach for N-linked glycan identification from MS/MS spectra by target-decoy strategy. *Comput. Biol. Chem.* **2018**, *74*, 391−398.

(33) Shipman, J. T.; Su, X.; Hua, D.; Desaire, H. DecoyDeveloper: An On-Demand, De Novo Decoy Glycopeptide Generator. *J. Proteome Res.* **2019**, *18*, 2896−2902.