A New Searching Strategy for the Identification of O-Linked Glycopeptides

Jiawei Mao,†*,‡‡ Xin You,†*,‡‡ Hongqiang Qin,†‡§ Chengye Wang,§ Liming Wang,§ and Mingliang Ye*†,‡‡

†CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Science, Dalian, 116023, China
‡University of Chinese Academy of Sciences, Beijing 100049, China
§Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, The Second Hospital of Dalian Medical University, Dalian, Liaoning 116023, China

Supporting Information

ABSTRACT: For the analysis of homogeneous post-translational modifications such as protein phosphorylation and acetylation, setting a variable modification on the specific residue(s) is applied to identify the modified peptides for database searching. However, this approach is often not applicable to identify intact mucin-type O-glycopeptides due to the high microheterogeneity of the glycosylation. Because there is virtually no carbohydrate-related tag on the peptide fragments after the O-glycopeptides are dissociated in HCD, we find it is unnecessary to set the variable mass tags on the Ser/Thr residues to identify the peptide sequences. In this study, we present a novel approach, termed as O-Search, for the interpretation of O-glycopeptide HCD spectra. Instead of setting the variable mass tags on the Ser/Thr residues, we set variable mass tags on the peptide level. The precursor mass of the MS/MS spectrum was deducted by every possible summed mass of O-glycan combinations on at most three S/T residues. All modifications were included in glycopeptide spectra matching. Compared with the conventional searching approach, O-Search yielded 96%, 86%, and 79% improvement in glycopeptide spectra matching, glycopeptide identification, and peptide sequence identification, respectively. It was demonstrated that O-Search enabled the consideration of more glycan structures and was able to analyze microheterogeneity of O-glycosylation.

Protein glycosylation is a highly heterogeneous post-translational modification (PTM), which mainly includes N- and O-linked glycosylation. It participates in various biological processes such as signal transduction, cell–cell adherence, and cell division.1,2 The aberration of glycosylation is associated with many diseases such as cancers and neurodegenerative diseases.3,4 It was reported that the glycan structures on one site varied among different samples and that the site-specific glycoform could be a potential biomarker for some diseases.5 Therefore, the global profiling of site-specific glycans has attracted increasing attention recently.5,6 Rapid progress has been seen for the analysis of site-specific N-glycoforms. A few excellent approaches were developed to decipher the MS/MS spectra of intact glycopeptides. For example, in 2014, Cheng et al. presented a twin spectra approach, by combining the spectra of intact and deglycosylated peptides, to identify intact N-glycopeptides which revealed 1769 site-specific N-glycans on 453 glycosites from HEK 293T cells.7 Instead of using two spectra, Eshghi et al. introduced a novel algorithm named “GPQuest” in 2015 for automatic identification of intact glycopeptides from its HCD (high-energy collision-induced dissociation) spectra based on spectral library matching.8 By using this algorithm, 769 unique intact N-glycopeptides were identified from 1008 glycan-containing spectra for LNCaP cells. Recently, Zeng et al.9 developed software (pGlyco) for the identification of intact N-glycopeptides with higher data interpretation efficiency, and 10 009 distinct site-specific N-glycans on 1988 glycosites from 955 glycoproteins in five mouse tissues were identified. Some of these approaches were also applied to identify intact O-glycopeptides with limited success.10 Currently, the identification of intact mucin-type O-glycopeptides is still challenging.

Received: September 13, 2018
Accepted: February 25, 2019
Published: February 25, 2019
For the analysis of homogeneous PTM such as protein phosphorylation or acetylation, setting a variable modification mass, e.g., 97.98 Da for phosphorylation and 42.01 Da for acetylation, on the specific residue(s) is applied to identify the modified peptides for database searching when the restricted search scheme is applied. However, this approach is often not applicable to intact mucin-type O-glycopeptides, as the search space will expand exponentially when multiple O-glycans are considered. The numbers of glycan compositions are given in brackets. Detailed information on glycan combinations and their composition are given in Table S1 (for 3 glycans) and S2 (for 12 glycans).

Figure 1. Expansion of the searching space in T-Search (traditional database search) and O-Search for O-glycopeptide identification. (A) A typical HCD spectrum of an O-glycopeptide; b and y ions were annotated without considering glycosylation. (B) The numbers of glycopeptide candidates in T-Search and O-Search after considering the glycan combinations. For O-search, different glycan combinations may result in the same glycan composition. The numbers of glycan compositions are given in brackets. Detailed information on glycan combinations and their composition are given in Table S1 (for 3 glycans) and S2 (for 12 glycans).
available in the Q Exactive machine. Instead, HCD is more popular for the analysis of glycoproteomes. Compared with CID in Orbitrap, HCD records both precursor ions and fragment ions with high mass accuracy, which makes the fragment ion matching more confident. More importantly, HCD provides more fragment ions especially b and y ions for peptide backbone sequencing due to the higher collision energy. A typical HCD spectrum of O-glycopeptide is shown in Figure 1A. Three types of fragment ions were observed: the fragments from the glycans (B ions or oxonium ions), the fragments from the peptide backbones (Y0, b/y ions), and the fragments derived from the loss of B ions from the intact glycopeptides (Y1, Y2...ions). Sufficient numbers of peptide fragments carrying PTM such as phosphorylation are required.

Figure 2. Schematics for O-Search to identify intact O-glycopeptides. (A) Remove nonglycopeptide spectra based on B ions and remove B ions in the glycopeptide spectrum. (B) Subtract the glycan mass from spectrum precursor mass to generate in silico deglycosylated spectrum and remove Y ions. (C) Database searching without setting variable glycan modification to identify peptide sequence for the glycopeptide with specified glycan. Spectra in B and C were annotated with the top matched peptide sequence (in blue).
to determine the modified sites in conventional database searching strategy. However, such fragment ions to determine the glycosites, i.e., the peptide fragments carrying the carbohydrate-related fragments, were not observed in the spectrum. According to our observation, these ions were rarely observed in the HCD spectra of other O-glycopeptides either. This means that the localization of glycosites to specific S/T residues is an impossible mission for most of the HCD data due to the lack of evidence. Because of the high diversity of the O-glycans, the conventional searching approach by setting multiple variable mass tags on the S/T residues to identify the peptide sequences from the HCD spectra of intact glycopeptides. Herein we present a novel approach called O-Search for the interpretation of O-glycopeptide HCD spectra. Instead of setting the variable mass tags on the S/T residue, we set variable mass tags on the peptide level. It was found that this method has much decreased search space and has excellent identification sensitivity in the identification of O-glycopeptides.

**MATERIALS AND METHODS**

**HILIC Enrichment of O-GalNAc Glycopeptides.** The tryptic peptides were dissolved in GlycoBuffer 2 (pH = 7.5), and PNGase F was added to release the N-linked glycans at 37 °C overnight. The de-N-glycosylated sample was desalted and subjected to O-GalNAc glycopeptide enrichment by using HILIC tip materials according to our previous reports. Briefly, the de-N-glycosylated digest was redissolved in the loading buffer (80% ACN/1% TFA), and 60 μL of solution equivalent to 50 μL of serum digest was pipetted into a HILIC tip with 5 mg of click-maltose materials. After capture, the HILIC tip was washed with 60 μL of loading buffer once and 20 μL of loading buffer twice. Finally, the captured O-glycopeptides were eluted with 100 μL of 30% ACN/1% TFA twice. The eluate was dried for high pH RPLC fractionation (Supporting Information) or direct RPLC-MS analysis.

**LC-MS/MS Analysis of O-GalNAc Glycopeptides.** The analysis of enriched human serum O-glycopeptides was performed using a Q-Exactive mass spectrometer (.raw) were converted into mzML format using Proteome Discoverer (version 1.4.0.0). Mascot (version 2.30) was used for the error-tolerant search. MSFagger (version 20181110) was used for open search. MS-GF+ (version 20180605) was used for the traditional restricted search, which was performed by setting the three most abundant O-linked glycans in human serum, NeuAc-Gal-GalNAc (~60%), NeuAc-Gal-(NeuAc-)GalNAc (~20%), and Gal-GalNAc (~10%), as variable modifications on amino acids Ser and Thr. The same glycans were applied to O-Search for comparison. ArMone uses a glycan database with 12 glycans, and O-Search analyzed these 12 glycans. The error-tolerant search and the open search were performed with similar settings except that these glycans were not set as variable modifications. The detailed parameter settings are provided in Supporting Information.

**RESULTS AND DISCUSSION**

**Workflow of O-Search.** This new searching scheme, termed as O-Search, is designed to identify O-linked glycopeptides from high-energy collision-induced dissociation (HCD) MS2 spectra. As shown in Figure 2, the analysis involves the following steps: (i) filtering the spectra to keep only glycopeptide spectra; (ii) generating in silico deglycosylated spectrum; (iii) peptide sequence identification. In this study, the presence of either of the two abundant B ions, NeuAcNexa→2H3O−CH2O− (m/z 138.05) and HexNAc+(m/z 204.09) (Figure S1B), was used to filter the raw data to keep the glycopeptide spectra for the first step. The removal of nonglycopeptide spectra by prefiltering would reduce the database search time and possibly produce more PSMs because nonglycopeptide spectra can only generate false positive glycopeptide spectrum matches and thus could increase the score threshold determined by the target-decoy analysis. In our data set, about 50% of acquired spectra were determined to be of glycopeptides for an LC-MS run. A typical spectrum for O-glycopeptide is shown in Figure 2A. The two abundant B ions as well as some low abundance B ions are presented. To facilitate the peptide sequence identification, all the B ions were removed from the spectrum. A much cleaner spectrum was generated as shown in Figure 2B.

O-GalNAc glycosylation often simultaneously occurs on multiple nearby S/T residues. For this reason, a tryptic glycopeptide could carry multiple O-glycans. In this study, a maximum of three glycosylation sites is considered for every peptide sequence. The masses for all possible combinations of glycans are generated. These masses are termed as glycan delta masses. For each combination, a new spectrum is generated by deducing the glycan delta mass from the precursor mass. In the meantime, the corresponding Y ions are removed from the spectrum. In this way, a glycopeptide spectrum was converted into a series of in silico deglycosylated spectra with different precursor masses. The number of in silico deglycosylated spectra corresponding to a glycopeptide spectrum depends on the number of O-glycan structures considered for searching. If only the three high abundant O-linked glycans in human sera, i.e., NeuAc-Gal-GalNAc (~60%), NeuAc-Gal-(NeuAc-)GalNAc (~20%), and Gal-GalNAc (~10%), were considered, then there are 19 types of combinations. Because some combinations yielded the same glycan composition, the 19 types of combinations yielded only 15 different spectra (Table S1). Thus, each glycopeptide spectrum will generate 15 in...
silico deglycosylated spectra. Many more spectra will be generated if more types of O-glycan structures are considered. For example, when the types of O-linked glycan structures increase to 12, the number of glycan combinations and delta masses increases to 454 and 184, respectively (Figure 1B). For this case, each raw spectrum will have 184 deglycosylated spectra.

All of the newly generated spectra were searched with the traditional database search method without setting any glycosylation variable modification. After searching, only the top matched peptide identification is kept for each spectrum. MS-GF+ is a universal database search tool optimized for a variety of spectral types.23 As an open source application, it can be easily integrated in our platform to perform database searching for peptide sequence identifications. Peptide spectrum matches (PSM) are filtered with minimum MS-GF:RawScore 15 and maximum MS-GF:EValue 0.01. For all of the in silico deglycosylated spectra of each spectrum, only the best PSM is kept. These PSMs are regarded as glycopeptide spectrum matches (GPSM). GPSM level false discovery rate (FDR) was maintained at <1% using the target-decoy approach. As shown in Figure 2C, the top matched peptide sequence for the spectrum was determined to be “TFVLSALQPSPTHSSNTQR”. The corresponding glycan delta mass for this peptide backbone identification is 656.23. Thus, the glycan composition was determined to be “N1H1F0S1” (HexNAc(1)Hex(1)Fuc(0)NeuAc(1)). This peptide has eight S/T residues. Because the peptide fragments carrying the carbohydrate-related fragments were rarely observed in the HCD spectrum, we are unable to localize the glycan to a specific S/T residue. Acquision of O-glycopeptide spectra by ETD may be helpful to localize glycosites. However, ETD is known to be poorly sensitive and is not equipped in Q-Exactive. It should be noted that only the glycan composition is determined in this strategy. It could be the combination of two or more glycans. In this study, the potential combinations were given in the “glycans” column in the identification result file. The glycan composition for the spectrum in Figure 2 should be only one glycan of NeuAc-Gal-GalNAc due to its small delta mass.

Performance Evaluation. This search strategy was then applied to analyze the human serum O-glycoproteome. The sample processing was the same as we reported previously. Briefly, the human serum was digested by trypsin and de-N-glycosylated by PNGase F. The O-glycopeptides were enriched by HILIC and then fractionated with high pH RPLC into ten fractions. Each fraction was analyzed by LC-MS/MS. For comparison, both O-Search and traditional database search (T-Search) methods by setting variable modifications were applied to process this data set. T-search with more than three glycans makes the total variable modification count higher than nine, which cannot be handled by most search engines. For this reason, only the three high abundant O-linked glycans in human sera, NeuAc-Gal-GalNAc (∼60%), NeuAc-Gal-(NeuAc-)GalNAc (∼20%), and Gal-GalNAc (∼10%), were set as the variable modifications on amino acids Ser and Thr in T-Search. For comparison, the same three glycans were considered in O-Search. The searching results are summarized in Figure S2. It was found that O-search outperform T-search at all three levels of GPSMs, glycopeptides, and peptide sequences (Figure S2A). O-Search yielded 96%, 86%, and 79% improvement in GPSMs, glycopeptides, and unique peptide sequences, respectively. It was found that 96% of the GPSMs, 94% of glycopeptides, and 88% of the peptide sequences identified by T-Search were covered by O-Search (Figure S2B). The majority of identifications obtained by T-Search can also be achieved by O-Search, which partly indicated the high confidence of the identifications achieved by O-Search. Though only 3 glycans were considered, 12 glycan compositions were observed on the glycopeptides (the glycan compositions identified with at least 10 counts are given in Figure 3A). This indicated the high heterogeneity of O-glycosylation.

Too many variable modifications make the search space of T-Search expand exponentially. The average number of S/T residues on O-glycopeptides identified in this study is about five (n = 5). Given that the number of glycan types is three (k = 3, the three abundant O-glycans in serum proteins), the

![Figure 3. Comparison of search results in O-Search by considering 3 and 12 glycan structures. (A) The distribution of the identified glycan compositions with at least 10 identification count, (B) the counts of and (C) the overlap of identified GPSMs, unique glycopeptides, and unique peptide backbones. The legends in B: “12 glycans” and “12 glycans (new)” means that the glycopeptides identified with the glycan compositions can be and cannot be generated by the combination of the three glycans.](https://doi.org/10.1021/acs.analchem.8b04184)
maximum number of glycosylations allowed on a peptide is three \((m = 3)\) and the search space is expanded by 375 fold (Figure 1B). For O-Search, there are only 19 different combinations with 15 masses. Thus, the searching space for O-Search increased by 15 fold, which is only 4% of T-Search. Smaller search space means faster search speed and high identification sensitivity, which makes O-Search a more practical tool than T-Search for the identification of O-glycopeptides. Because O-glycan chains are lost in most fragments, it is not necessary to generate theoretical MS/MS spectra of peptides with PTM combinations. T-search (represents existing restricted searching tools) does not perform well because it assumes that modifications remain on fragments in tandem mass spectra, which is not true for O-glycopeptides. This inappropriate assumption leads to the poor peptide-spectrum match quality and less identification efficiency of T-Search. Indeed, O-Search demonstrated higher match scores for 90% of the GPSMs identified by both T-Search and O-Search (Figure S3). When the considered glycans were extended to 12 glycans (Figure S1), the search space increased by 18 780 fold for T-search while only 184 fold for O-Search (Figure 1B). For this scenario, T-Search is unable to identify the O-glycopeptides while O-Search still works very well. As shown in Figure 3B, O-Search considering 12 glycans yielded the identification of 1115 PSMs, 248 glycopeptides, and 100 peptide backbones. Compared with the search considering three glycans, 52% more for GPSMs, 66% more for glycopeptides, and 30% more for peptide backbones were identified. Totally, 77 different glycan compositions were identified. Among them, 24 were identified with at least 10 counts (Figure 3A). Among these 24 relative abundant glycan compositions, 14 can only be identified by considering 12 glycans. Some glycan compositions were quite abundant based on the spectra counts. For example, N2H0F0S2, which was composed of two NeuAc(1)HexNAc(1), accounted for 8% of all GPSM counts. Those GPSMs cannot be achieved by considering the three glycans. It can be seen from Figure 3B and 3C that the improved coverage was achieved only because the additional glycan compositions were considered. It can also be observed that some glycopeptides could only be identified from the search with three glycans. We compared the q-value of the 75 PSMs missed by O-Search with 12 glycan types and found that they have relatively higher q-values and so they did not pass the criteria with FDR < 0.01 (Figure S4). This is because the searching space for 12 glycans increased about 10 times which lead to poor sensitivity. Clearly the number of glycans considered for searching should be carefully considered. Too many glycans will lead to an increase in random matches, and too few glycans will miss the identification of some glycopeptides carrying other glycans. In this study, the 12 glycan searching resulted in the missing of 27 glycopeptides while 126 glycopeptides were newly identified (Figure 3C). Clearly the searching with 12 glycans revealed more site-specific O-glycans.

In the silico deglycosylation strategy (ArMone) we developed previously is able to identify intact O-glycopeptides when multiple O-glycans are considered. Open search and error-tolerant search are also unrestricted by the number of modifications. We compared their performance with O-Search by considering the 12 glycans. O-Search identified 99.5% more GPSMs, 79.7% more distinct glycopeptides, and 53.8% more distinct peptide sequences than the second-best strategy, i.e., ArMone (Figure S5A). O-Search covers the majority of identifications achieved by the other three strategies (Figure S5B, S5C, S5D). Evidently, the identifications achieved by O-Search are significantly more comprehensive. In ArMone, only the MS2 spectra with Y0 ion (peptide backbone ion) were in silico deglycosylated for peptide identification. The spectrum shown in Figure S6 is absent of Y0 ion. This spectrum yielded the identification of a glycopeptide, TVVQPSVGAAGPVVPPCPGR (HexNAc(2)NeuAc(1)), by using O-Search. However, it did not yield any identification in ArMone. The fact that ArMone is unable to identify O-glycopeptides from the spectra lacking Y0 ions may partly explain its poor performance. Open search uses a fragment ion matching scheme to match peptide sequences which is similar to O-Search, but it only identified less than 20% of GPSMs achieved by O-Search. This is mainly because open search checks a much wider precursor mass tolerance window, 0–1000 Da here, while O-Search only checks some discrete mass intervals. This makes the search space of open search hundreds of thousands of times larger than O-Search and results in the lower identification sensitivity of open search (Figure S5). The error-tolerant search identified the least number of GPSMs, and many reasons could lead to its poor performance, such as dependence on the presence of unmodified peptides, search space expansion caused by the checking of all modifications, only one modification allowed for each peptide, etc. Clearly, O-Search performs better than these strategies for the identification of O-glycopeptides.

By using our O-search strategy when considering 12 glycans, 248 intact O-glycopeptides with 100 unique peptide backbones corresponding to 63 O-glycoproteins were identified from human serum. Among these proteins, 39 proteins (61.9%) were previously reported to be O-GalNAcylated.22-24 As shown in Figure S8, identified O-glycopeptides with two or more PSM were mostly (32/36, 88.9%) known to be modified by mucin-type O-glycosylation. These results proved the reliability of our O-search algorithm in characterizing high-abundant O-glycoproteins. For the glycoproteins identified by one peptide, 38.1% (8/21) of them were known O-GalNAcylated proteins. This indicated that our approach was also able to identify low-abundant O-glycoproteins in biological samples. The excellent performance of this method can mainly be attributed to the improved fragment ion matches without considering the ions carrying glycans. As illustrated in Figure S9, many abundant peaks cannot be annotated if the glycan is considered to be carried on the peptide fragments while almost all abundant ions are annotated when no glycan is considered.

The N-linked glycans may not be completely removed by PNGase F, which could result in false assignments in O-Search. The common core sequence of eukaryotic N-glycans, Man3-3(Manα1–6)Manβ1–4GlcNAcβ1–4GlcNAcβ1–Asn-X-Ser/Thr, leads to a characteristic Y-ion pattern in the N-glycopeptide spectrum, which can help us identify the N-glycopeptide spectra. In this study, 100 unique peptide sequences were identified by O-Search. Among them, only 20 contain N-glycopeptide sequence motifs (Asn-X-Ser/Thr, X ≠ Pro). We checked the 143 spectra of the 20 peptide sequences with N-glycopeptide sequence motif and found that only 6 of the spectra corresponding to 2 peptide sequences were identified to hold Y-ion patterns (Figure S7). The above results indicated that the residual N-glycosylation has no serious effect on the identification of O-glycopeptides in this study.
Microheterogeneity of O-Glycosylation. Our approach is able to reveal the microheterogeneity of O-glycosylation at the proteome level. For every identified O-glycopeptide, we reported the attached glycan composition. Figure 4A shows the number of unique glycan compositions carried by each peptide backbone. Many peptide backbones carried multiple glycan compositions. The top one (716IEETTMTTQTPAPIQAPSAILPLPGQSVER745), carrying 22 different glycan compositions, came from ITIH4_HUMAN. ITIH4_HUMAN is a well-known O-glycoprotein, and the residues T719, T720, and T722 are annotated as O-linked glycosites in UniProt. These glycosites are on this identified peptide. Another peptide with a missed trypsin cleavage site, 711VMNMKIEETTMTTQTPAPIQAPSAILPLPGQSVER745, also includes these glycosites and was determined to carry nine different glycan compositions. The microheterogeneity of O-linked glycosylation could be well exemplified with ITIH4_HUMAN (Figure 4B). In addition to the range from 711 to 745, another range, from 690 to 710, was also determined to carry 6 O-glycan compositions. The peptide backbone in this range (690LAILPASAPPATSNPDPAVSR710) was reported to be O-glycosylated in a recent study analyzing the ITIH4 purified from human serum, indicating the high confidence of the identification. The left two peptides, i.e., 689RLAILPASAPPATSNPDPAVSR710 and 513LPTQNIITFQTESSVAAE-QEAFQSPK537, on this protein were identified to carry only one glycan composition. The peptide in the range of S13 to S37 has not been reported to be O-glycosylated. This intact glycopeptide was identified with a spectra count of only one. This region could be O-glycosylated with very low stoichiometry, and the identification was missed in other studies. If we consider it a false positive identification, the FPR at GPSM would be 0.5%, as the total count for all intact glycopeptides of ITIH4_HUMAN was 215. This also means that the identifications achieved in this study are of high confidence.

CONCLUSION
In this study, we presented a new search scheme called O-Search for the interpretation of O-linked glycopeptide HCD spectra. It was applied to analyze the O-glycoproteome of serum and was found to significantly outperform the conventional searching scheme. Because no variable glycan modifications were set during database searching, the searching space was significantly reduced and many more glycan structures could be considered. This makes this approach a powerful tool to reveal the microheterogeneity of O-glycosylation.

ASSOCIATED CONTENT
*
Supporting Information
The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b04184.

Figures S1–S9, describing oxonium ions, annotated peptide spectra, and identification results. Tables S1–S7, describing glycan combinations and identification results (PDF)
Table S2: Glycan combinations for considering 12 glycans (XLSX)
Table S3: Identification results of ArMone (XLSX)
Table S4: Identification results of O-Search with 3 glycans (XLSX)
Table S5: Identification results of O-Search with 12 glycans (XLSX)
Table S6: Identification results of error-tolerant search by Mascot (XLSX)
Table S7: Identification results of open search by MSFragger (XLSX)

AUTHOR INFORMATION

Corresponding Author
*E-mail: mingliang@dicp.ac.cn.

ORCID
Hongqiang Qin: 0000-0002-7508-0872
Mingliang Ye: 0000-0002-5872-9326

Author Contributions
Jiawei Mao and Xin You contributed equally to this work.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS
The authors acknowledge funds from the China State Key Basic Research Program Grants (2018YFC0910302, 2016YFA0501402, 2017YFA0505003), the National Natural Science Foundation of China (21535008, 81430072, 21775146), and the innovation program (DICP TMSR201601) of science and research from the DICP,
CAS. M.Y. is a recipient of the National Science Fund of China for Distinguished Young Scholars (21525524).

REFERENCES