analytical chemistry

Article

pubs.acs.org/ac

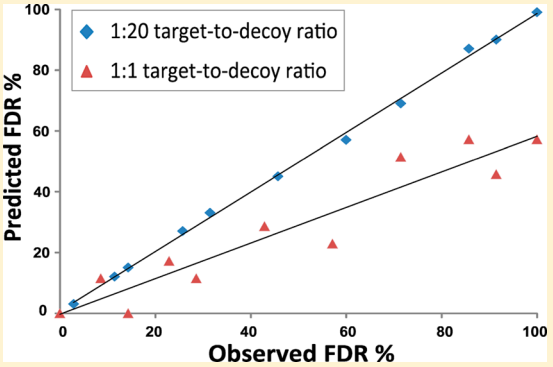
# New Glycoproteomics Software, GlycoPep Evaluator, Generates Decoy Glycopeptides de Novo and Enables Accurate False Discovery Rate Analysis for Small Data Sets

Zhikai Zhu, Xiaomeng Su, Eden P. Go, and Heather Desaire*

Department of Chemistry, University of Kansas, Lawrence, Kansas 66047, United States

Supporting Information

**ABSTRACT:** Glycoproteins are biologically significant large molecules that participate in numerous cellular activities. In order to obtain site-specific protein glycosylation information, intact glycopeptides, with the glycan attached to the peptide sequence, are characterized by tandem mass spectrometry (MS/MS) methods such as collision-induced dissociation (CID) and electron transfer dissociation (ETD). While several emerging automated tools are developed, no consensus is present in the field about the best way to determine the reliability of the tools and/or provide the false discovery rate (FDR). A common approach to calculate FDRs for glycopeptide analysis, adopted from the target-decoy strategy in proteomics, employs a decoy database that is created based on the target protein sequence database. Nonetheless, this approach is not optimal in measuring the confidence of *N*-linked glycopeptide matches, because the glycopeptide data set is considerably smaller compared to that of peptides, and the requirement of a consensus sequence for *N*-glycosylation further limits the number of possible decoy glycopeptides tested in a database search. To address the need to accurately determine FDRs for automated glycopeptide assignments, we developed GlycoPep Evaluator (GPE), a tool that helps to measure FDRs in identifying glycopeptides without using a decoy database. GPE generates decoy glycopeptides de novo for every target glycopeptide, in a 1:20 target-to-decoy ratio. The decoys, along with target glycopeptides, are scored against the ETD data, from which FDRs can be calculated accurately based on the number of decoy matches and the ratio of the number of targets to decoys, for small data sets. GPE is freely accessible for download and can work with any search engine that interprets ETD data of *N*-linked glycopeptides. The software is provided at https://desairegroup.ku.edu/research.




Glycosylation is commonly considered the most extensive post-translational modification on proteins, and it is estimated that 20%−50% of all proteins are glycoproteins.[1,2] Glycosylation is known to impact protein folding and function;[3,4] the interaction between proteins and glycans is a main route for cellular communications and signaling.[5−7] In addition, changes in glycosylation pattern on certain proteins are closely related to the pathogenesis of diseases.[8,9] Therefore, protein glycosylation analysis is a vital step toward understanding the role that carbohydrates play in various biological events.

One common method of characterizing the glycosylation on proteins is to digest the protein and to analyze the resulting glycopeptides. This strategy allows researchers to correlate the glycans to their attachment sites in the protein(s).[10−12] In glycopeptide analysis, the correct glycopeptide compositions usually cannot be determined by high resolution MS data alone, and MS/MS data are needed for confident glycopeptide assignments.[13] In order to accelerate the analysis workflow for high-throughput glycopeptide identifications, an increasing number of bioinformatics tools are developed to analyze MS/MS data of glycopeptides.[14−20] Strum et al. presented a program called GlycoPeptide Finder that can interpret CID data of *N*- and *O*-linked glycopeptides generated from nonspecific proteolysis.[21] A computational framework was developed to implement a software tool called GlycoFragwork, which is capable of scoring *N*-linked glycopeptide MS/MS data from multiple fragmentation modes.[22] We recently introduced two web-based utilities, GlycoPep grader[23] and GlycoPep Detector,[24] to determine the most likely *N*-linked glycopeptide compositions by scoring the CID and ETD data against each of the possible glycopeptide candidates. In all the applications described above, the glycopeptide analysis tool returns a best glycopeptide match for each MS/MS spectrum by selecting the candidate that receives the highest score under a certain scoring algorithm. Although these matches are very helpful in guiding the user, the top match is sometimes incorrect.

While automated analysis tools are helpful for glycopeptide analysis, users need to know the likelihood that the automated matches are correct. Therefore, it is important for any tool to

**Received:** June 12, 2014
**Accepted:** August 19, 2014
**Published:** August 19, 2014

   dx.doi.org/10.1021/ac502176n | *Anal. Chem.* 2014, 86, 9212−9219

provide users with a reliable false discovery rate (FDR), which is the measure of probability that a match is correct, based on the program's performance in analyzing the entire data set.[25−28] The concept of calculating an FDR has been well established by the proteomics community, and to determine the FDR value in proteomics, a composite database is generated by combining the target protein sequence database and a decoy sequence database. The decoy database is nonsensical and created based on the target database such that they contain an equivalent number of peptide sequences, which is often accomplished by reversing the protein sequences in the target database.[26−30] Subsequently, the MS/MS data are scored against the composite database, and the numbers of matches made against the target and decoy sequences are used to calculate FDR. Following the assumption that the distribution of incorrect matches to target sequences is the same as that of matches to decoy sequences, the number of false positive identifications, which directly translates to FDR, can be calculated by doubling the number of decoy matches. This target-decoy approach is simple and works well for peptide identifications based on large scale proteomics data.[31−33]

Most of the currently available glycopeptide analysis tools do not have the capability to calculate FDRs for glycopeptide assignments, and for those that are enabled with this functionality, the target-decoy approach is adopted to estimate FDRs in glycopeptide identifications, where an equal amount of decoy glycopeptides are generated on the basis of the target glycoprotein sequences to comprise the decoy database.[21,22] However, in a glycoproteomics experiment, the number of CID or ETD spectra scored is considerably smaller than the number of spectra scored in a proteomics experiment. This is expected since glycoproteomics experiments are often conducted on a single protein, not thousands of proteins. Even when the entire proteome is evaluated for glycopeptides, the number of CID or ETD spectra that are verified to be from glycopeptides is generally much less than 1000. As a result, using the conventional approach for calculating FDRs, the distribution of decoy glycopeptide matches may not accurately reflect that of incorrect matches to target glycopeptides because the collected glycopeptide data set is not large enough.[21,34,35] Furthermore, for *N*-linked glycopeptides, a consensus sequence of N-X-S/T (X can be any amino acid except proline) must be present, which further limits the number of possible decoy glycopeptides being tested. All these factors lead to inaccurate FDRs when the target-decoy approach is applied to small to moderate size glycoproteomics data sets.

In this work, we present a new method to determine FDRs with high accuracy for *N*-linked glycopeptide identifications based on ETD data. Instead of creating a decoy database of the same size as the target database, we developed a tool called GlycoPep Evaluator (GPE) to generate decoy glycopeptides de novo for every target glycopeptide, in a 1:20 target-to-decoy ratio. The decoys are made under specific rules so that they contain the consensus sequence for *N*-linked glycosylation, while they have distinct glycopeptide sequences and glycosylation sites. To determine the FDR, all the generated decoys are scored against the ETD data along with target glycopeptides, and the FDR is calculated accurately based on the number of decoy glycopeptide matches and the relative amount of targets to decoys. GPE is freely available for download and can be used in conjunction with any scoring schemes for assessing ETD data of glycopeptides. Please visit https://desairegroup.ku.edu/research for a copy of the software.

## EXPERIMENTAL SECTION

**Samples and Reagents.** Bovine fetuin, RNase B, and human serum proteins (IgG, AGP, transferrin) were obtained from Sigma-Aldrich (St. Louis, MO). The HIV envelope protein, C.97ZA012 gp140, was provided by the Duke Human Vaccine Research Institute (Durham, NC).[36] Sequencing grade trypsin was purchased from Promega (Madison, WI). All chemical reagents used were either of analytical grade or better.

**Protease Digestion.** Glycoproteins of 72−100 $\mu$g were dissolved in 100 mM Tris buffer at pH 8 with a concentration of 2.4−3.3 $\mu$g/$\mu$L. Samples were denatured by addition of urea so that the final urea concentration was 6 M, followed by addition of 5 mM tris(2-carboxyethyl)-phosphine (TCEP) solution to reduce the disulfide bonds (the molar ratio of TCEP to disulfide bond was kept at 6:1), and 10 mM iodoacetamide (IAM) was subsequently added to alkylate the free thiol groups using a molar ratio of 8:1. The reaction was left to proceed for 1 h at room temperature in the dark. Dithiothreitol (DTT) solution was then added to a final concentration of 10 mM to quench the alkylation reaction. Prior to enzymatic digestion, the urea concentration was decreased to 1 M by diluting the samples with Tris buffer. Subsequently, trypsin was added at a 1:30 enzyme-to-protein ratio, followed by 18 h incubation of the samples at 37 °C. Finally, trypsin digestion was stopped by adding 1 $\mu$L of acetic acid for every 100 $\mu$L of glycoprotein solution. The prepared samples were stored at −20 °C before subjected to LC/MS analysis.

**LC/MS Analysis.** Digested glycoprotein samples were analyzed using a Waters Acquity Ultra Performance Liquid Chromatography system (Milford, MA) coupled to a LTQ Velos linear ion trap mass spectrometer (Thermo Scientific, San Jose, CA). For each run, 5 $\mu$L of a sample was injected onto a capillary $C_{18}$ column (300 $\mu$m i.d. × 5 cm, 100 Å, Micro-Tech Scientific, Vista, CA). Two mobile phases were employed for separation: solvent A consists of 99.9% $H_2O$ plus 0.1% formic acid, and solvent B consists of 99.9% acetonitrile with 0.1% formic acid. The LC separation gradient was as follows: 2% solvent B for 5 min, followed by a linear increase to 40% B in 50 min, and a ramp to 90% B in 10 min.[37,38] The column was kept at 90% solvent B for an additional 10 min and then re-equilibrated at 2% B for 10 min. The mass spectrometer was operated in the positive ion mode, with the ESI source voltage at 3 kV and the capillary temperature set at 200 °C. For the data-dependent acquisition, CID and ETD spectra were collected by selecting the five most intense peaks in the full scan MS ($m/z$ 500−2000) and the precursor ions were fragmented in either CID or ETD mode. In the MS/MS settings, automatic gain control (AGC) function was enabled with a target value of $2 \times 10^4$ for the ion trap; the fluoranthene anions, employed for ETD fragmentation, was set at a AGC target value of $2 \times 10^5$. The reaction time between anions and cations in ETD was set at 90 ms, and the supplemental activation was turned on for ETD so that precursor ions and charge-reduced species could undergo further dissociation. For CID, the normalized collision energy was set at 30%, with activation time of 10 ms.

**Glycopeptide MS/MS Data Set.** In this study, MS/MS data were collected on glycoproteins that have been previously characterized in the literature.[36,39−42] In silico trypsin digestion


dx.doi.org/10.1021/ac502176n | *Anal. Chem.* 2014, 86, 9212−9219

was performed on the glycoprotein sequences with up to 2 missed cleavages allowed, and carbamidomethylation was set as a fixed modification on cysteine residues. Theoretical monoisotopic masses of potential *N*-linked glycopeptides were calculated by adding the site-specific glycan masses to the masses of the corresponding peptides that contain the glycosylation sites. The theoretical $m/z$ values of these glycopeptides were then computed and searched against the ETD data to see whether precursor ions of these $m/z$ values were selected for ETD. Manual analysis was then performed on every identified ETD spectrum that may come from potential glycopeptides. If a match was found, CID data were employed to further confirm the glycopeptide assignment. In this way, a glycopeptide ETD data set with known glycopeptide compositions was built that includes glycopeptides of diverse peptide sequences and varying glycan types.

**Decoy and Target Candidates Generation.** For this study, all of the glycopeptide assignments were known. However, to demonstrate our approach, we simulated a case where the identity of the glycopeptide was not known and the user had to choose between multiple feasible candidates. Therefore, we needed mock candidates and decoys to score against each spectrum. GlycoPep Evaluator (GPE) was used to generate 20 decoys per candidate. The correct "candidate" for each spectrum is known, and the additional mock candidates were generated using GlycoMod.[42] To generate the mock candidates, sequences of the studied glycoproteins were entered into GlycoMod, along with a polypeptide sequence, Titin, which contains 50 000 amino acid residues. The mock candidates contain the consensus motif of N-X-S/T, and their glycan compositions are biologically relevant. As a result, multiple glycopeptide compositions were produced by GlycoMod for every glycopeptide peak that was subjected to ETD (with a mass tolerance of 200 ppm), and a selection of the glycopeptides were entered into GPE as (mock) target glycopeptide candidates. Typically, five candidate glycopeptides were entered, where one of the candidates was the true glycopeptide. For each target glycopeptide, GPE is used to generate 20 decoy glycopeptides of isobaric masses, and these decoys can be used for evaluating the false discovery rate (FDR) in automated assignment of glycopeptides by a search engine. (GPE includes functionality to generate any number of decoys, but 20 were used here.)

**Scoring of Decoy and Target Candidates.** GPE is a freely available software tool that we developed to assist in determining FDRs in glycopeptide analysis. The function to generate decoy glycopeptides is the main innovation of this tool, and the algorithm used to generate the decoys is described in detail in the Results section. GPE also incorporates an ETD algorithm that we described previously,[24] and it can score each target and decoy candidate against the ETD spectrum in an automated manner. The software may be used as a standalone program simply for generating decoys, or it can be used to score the input decoys and targets using the embedded scoring tool. In order to use the scoring functionality of GPE, the user needs to upload a raw ETD data, specify the MS/MS scan range and the ion types being scored, and submit the target and decoy candidates for scoring. GPE then generates the result page where the candidates are ranked by the scores that they are assigned.

**FDR Study Using GPE.** GPE was used to score a set of ETD spectra from 77 different glycopeptides, which had been manually assigned, as described above. The software generated decoy glycopeptides for all the input target glycopeptides, and a target or a decoy match was made depending on whether a target or a decoy candidate received the highest score. Using the number of decoy matches made by GPE in assessing the glycopeptide data set and the target-to-decoy ratio (1:20 in our study), the FDRs in glycopeptide analysis could be calculated.

## RESULTS AND DISCUSSION

**Overview of GlycoPep Evaluator.** GlycoPep Evaluator (GPE) is a freely downloadable software tool that can be used to generate decoy glycopeptides for false discovery rate analysis. GPE is available for download at https://desairegroup.ku.edu/research. It has incorporated functionality to score all the targets and decoys against imported spectra using a previously published scoring algorithm.[24] GPE was written in Java and developed with Java Development Kit 7 (JDK 7). The program has been tested to perform successfully under Windows and Linux systems, and Java Runtime Environment 7 (JRE 7) is recommended to be installed prior to running GPE.

The graphical user interface (GUI) of GPE is shown in Figure 1A. To generate decoy glycopeptides, the user needs to enter the target glycopeptide sequence and to specify the *N*-glycosylation site location by entering the Glycosylated Asn Index (if a default value of 0 is input, the software will automatically locate the first Asn that meets the N-X-S/T sequon). Cysteine modifications can be selected by the user as indicated in the GUI; if there is an additional modification on any amino acid residue, the user can specify the location and the mass of the modification as needed. For the glycan portion, the user can either type in the number of each monosaccharide unit (Hex, HexNAc, Neu5Ac, etc.) that constitutes the glycan or input the glycan mass, as shown in Figure 1A. Other parameters that are necessary to generate decoys include the precursor ion's $m/z$ and charge state, mass tolerance (in ppm), number of maximum missed cleavages, peptide variation (in Da, see discussion below) and the number of decoys per target. The mass tolerance is the mass range that the monoisotopic mass of a decoy glycopeptide, as generated by GPE, is allowed to deviate from the precursor ion's mass (as calculated by the precursor ion's $m/z$ and charge state). The peptide variation, on the other hand, is the mass range that the peptide portion of the decoy (calculated by subtracting the glycan mass from the monoisotopic mass) is allowed to differ from that of the peptide in the target glycopeptide. In our experiments, the mass tolerance for decoys was set at 20 ppm, maximum missed cleavage number was set to 2, peptide variation was set at 200 Da and the number of decoys per target was set to 20. Currently, the tool specifically generates tryptic peptides. If sufficient interest warrants future development, other options for peptide generation could be included.

Once the required parameters are submitted to generate decoy glycopeptides, GPE will present the result page where 20 output decoys are listed, as exemplified in Figure 1B. Several requirements are met by GPE in producing the decoy glycopeptide candidates: First, the decoy ends with either Arg or Lys on its C-terminus; second, the missed cleavages on the decoy sequence must not exceed the number of maximum missed cleavages specified by the user; third, the decoy contains a consensus sequence, Asn-X-Ser/Thr (X is any random amino acid, excluding proline), with the Asn being the glycosylation site; fourth, the peptide portion of the decoy has a mass that is within a user-specified range (termed "peptide variation") from the peptide mass of the target glycopeptide; finally, the glycan


dx.doi.org/10.1021/ac502176n | *Anal. Chem.* 2014, 86, 9212−9219

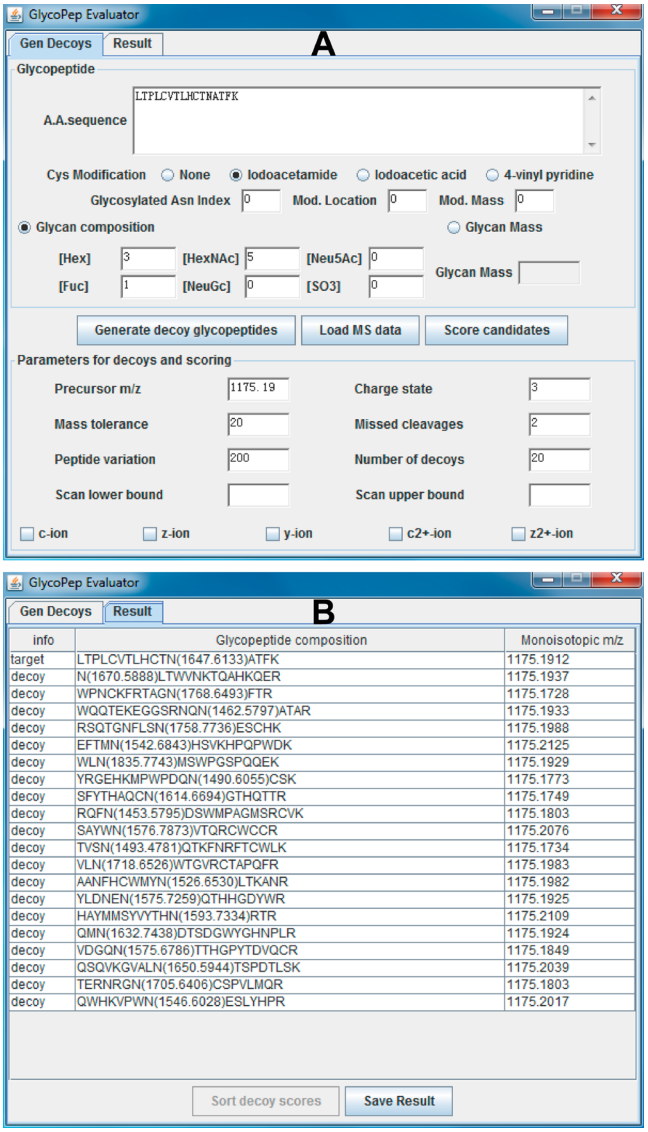



**Figure 1.** (A) Graphical user interface (GUI) of the GlycoPep Evaluator (GPE) program. (B) The result of decoy generation completed by GPE that contains the input target glycopeptide as well as 20 decoy glycopeptide sequences generated by the program.

portion of the decoy is assigned a mass that makes the $m/z$ of the entire decoy within the user-specified mass tolerance of the precursor ion's $m/z$, and the glycan mass value is appended to the glycosylated Asn as a modification of mass in the output of the decoy glycopeptide (Figure 1B).

Following these rules, the generated decoy glycopeptide can closely mimic the target glycopeptide in terms of the glycosylation site, protease specificity, and the approximate peptide length. On the other hand, 20 decoy glycopeptides of distinct sequences and varying glycan locations are produced for every single target glycopeptide, as demonstrated in Figure 1B, thus providing a sufficient number of decoy candidates that can compete with the target glycopeptides in the scoring by a software tool.

**False Discovery Rate Analysis.** The false discovery rate (FDR) is, by definition, the percentage of accepted peptide-spectral matches that are incorrect.[28] When decoys are included in database searching, the incorrect matches are comprised of a proportion of the target matches as well as all the decoy matches. The latter are used to estimate the number of target matches that are incorrect. As such, FDR is calculated by the following equation:

$$\mathrm{FDR} = \frac{N_{ic} + N_{d}}{\text{total assignments}} \tag{1}$$

In the equation, $N_{ic}$ is the number of incorrect assignments made to target candidates and $N_{d}$ is the number of decoy assignments.

Because both the incorrect target matches and the decoy matches are made at random, the number of hits for incorrect target assignments or decoy assignments is proportional to the number of the corresponding target or decoy candidates scored by a program. Consequently, the ratio of the number of incorrect target assignments to decoy assignments is equal to the ratio of target candidates to decoy candidates in quantity:

$$\frac{N_{ic}}{N_{d}} = \frac{\text{number of targets}}{\text{number of decoys}} \tag{2}$$

When eqs 1 and 2 are combined, the FDR is determined by eq 3:

$$\mathrm{FDR} = \frac{N_{d}}{\text{total assignments}}\left(1 + \frac{\text{number of targets}}{\text{number of decoys}}\right) \tag{3}$$

In a conventional workflow, since an equal number of decoy sequences are scored along with target sequences, $N_{ic}/N_{d}$ is 1. Therefore, according to eq 3, the FDR is calculated by doubling the number of decoy matches divided by the number of total assignments. In our method, however, the target-to-decoy ratio is 1:20 rather than 1:1 because 20 decoy candidates are generated and scored for each target, thus $N_{ic}/N_{d}$ is 0.05. Accordingly, FDR is determined by eq 4:

$$\mathrm{FDR} = \frac{N_{d}}{\text{total assignments}}1.05 \tag{4}$$

Consequently, using our method in which 20 decoy glycopeptides are created and tested for every target glycopeptide composition, the FDR can be measured accurately based on the number of decoy matches and the number of total accepted assignments, as formulated in eq 4.

**Target and Decoy Glycopeptides Analysis.** Apart from generating decoy glycopeptide candidates de novo, GPE was also implemented with an algorithm that we developed to process and score ETD data of *N*-linked glycopeptides.[24] After a list of decoy candidates are generated by GPE, the user can load raw ETD data to the program and specify the MS/MS scan range; GPE can score all the decoy candidates as well as the target glycopeptide compositions against the input MS/MS data. For every glycopeptide composition, GPE evaluates the match of different ion series (c, z, and y-ions) to the processed ETD data and assigns a final score to each candidate, as described in the algorithm published with ref 24. The decoy glycopeptides can then be sorted from high score to low and be compared with the scores of target glycopeptides.

To demonstrate the functionality of GPE, we present, below, a CID and ETD spectrum of a known glycopeptide and show how the GPE would process the ETD data, score the spectrum, and then additionally calculate scores for decoy assignments. Figure 2A is the ETD data of a glycopeptide from HIV gp140 that has a composition of DGGED***N***KTEEIFRPGGGNMK + [Hex]3[HexNAc]4[Fuc]1 (where ***N*** is the glycosylation site).


dx.doi.org/10.1021/ac502176n | *Anal. Chem.* 2014, 86, 9212−9219

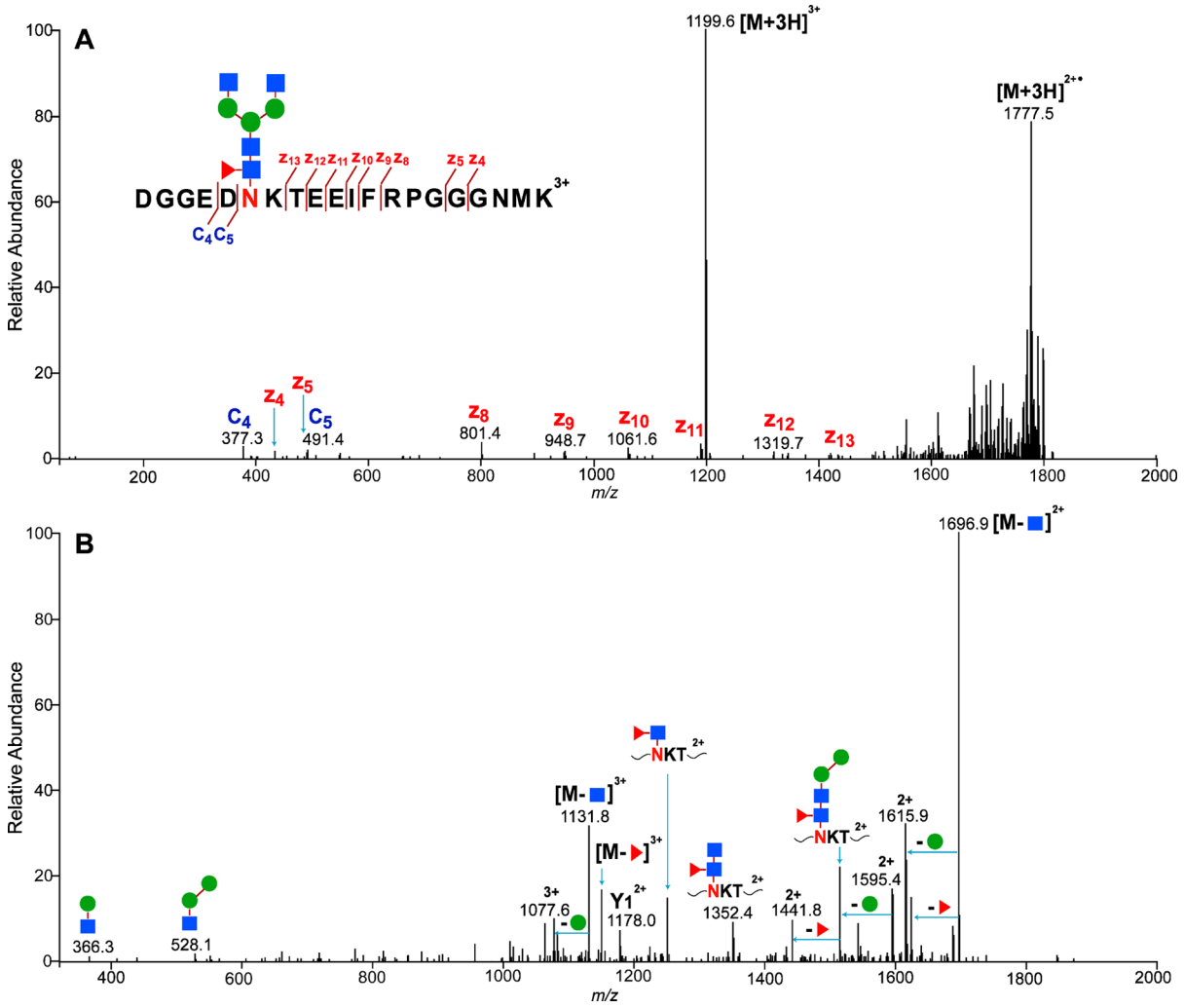



**Figure 2.** (A) ETD-MS/MS data of a HIV gp140 glycopeptide that has a core-fucosylated biantennary complex-type glycan as shown in the figure. The peptide backbone fragment ions (c- and z-ions) are labeled. (B) CID data of the same glycopeptide in (A). Extensive dissociation at the glycan portion is observed in CID; product ions containing partially cleaved glycans and intact peptide sequences are present in the data. This figure is an example of how spectra were manually assigned, prior to testing of GPE. Please note: generally the glycan composition, but not the structure, is confirmed.

In the ETD spectrum, c-ions ($c_4$-$c_5$) and z-ions ($z_8$-$z_{13}$) are observed that can be used to determine the glycopeptide's sequence, as shown in the figure. Additionally, the CID data in Figure 2B further confirms that the precursor ion is a glycopeptide peak, because glycan oxonium ions are present at $m/z$ 366 and 528. Moreover, by assigning monosaccharide losses, including losses of Hex, HexNAc and glycan dissociation patterns in CID, the glycan portion of the glycopeptide can be deduced to be [Hex]3[HexNAc]4[Fuc]1. It is noteworthy that, although CID data are utilized to verify glycopeptide assignments, in our method, we did not implement CID fragmentation rules in the scoring function, and only ETD data should be submitted to GPE for appropriate FDR analysis.

To demonstrate that the glycopeptide composition described above can be correctly assigned by GPE, the true glycopeptide composition, along with four isobaric glycopeptide "mock" candidates, were entered into GPE as potential target glycopeptide candidates. GPE then generated 20 decoy glycopeptides per target. The ETD data were subsequently submitted to the software, and all the candidates (including decoys) were scored by GPE. A total of 100 decoy glycopeptide compositions were created by GPE for the 5 target glycopeptides, and each decoy has its distinct sequence and glycosylation site. The true glycopeptide composition, its 20 decoys, and the associated scores are shown in Figure 3; the remaining 4 targets, their 80 decoys, and their scores are shown in the Supporting Information, Table 1. The correct glycopeptide composition, labeled as target in Figure 3, receives the highest score of 61.7, which is significantly higher than the score of any other candidate, including the other 4 target and 100 decoy glycopeptides. By contrast, none of the other 4 input target glycopeptides (which are incorrect candidates but still considered "targets", for the purposes of this demonstration), outscore the best decoy glycopeptide sequences generated by GPE. While at least one of the 20 decoys in each of these sets outscore the falsely generated "target" candidates, the overall highest scoring decoy, with a score of 17.8, does not outscore the true assignment. (Additional data are shown in Supplementary Table 1.) Therefore, the first glycopeptide candidate, which is also the manually verified correct assignment, is assigned to the ETD data by GPE, even when four other incorrect candidates and 100 decoys are scored in parallel. This


dx.doi.org/10.1021/ac502176n | Anal. Chem. 2014, 86, 9212−9219

| Name | Composition | Mono *m/z* | Score |
|---|---|---|---|
| target | DGGEDN(1444.5339)KTEEIFRPGGGNMK | 1199.1752 | 61.6754 |
| decoy 1 | N(1588.5408)VSCKKPVWCMPSGGGSR | 1199.1637 | 8.8906 |
| decoy 2 | YDN(1588.6667)RSMRGMFGFTHCK | 1199.1865 | 8.7011 |
| decoy 3 | AKYPTGN(1380.4808)DTFENESDVKGSR | 1199.1739 | 6.2219 |
| decoy 4 | YFWCPKRWQN(1667.5246)LSSR | 1199.1592 | 5.9449 |
| decoy 5 | LN(1437.3562)LTWSKHTHNPRLPTDK | 1199.1741 | 5.4633 |
| decoy 6 | VFVLVGNATN(1413.4413)MSESWCDPR | 1199.1538 | 4.5258 |
| decoy 7 | N(1158.3110)MTKWPNEHWRNGMLHMVR | 1199.1614 | 4.3031 |
| decoy 8 | PEADMQGSREPN(1461.4535)ATFWAVK | 1199.1568 | 4.1652 |
| decoy 9 | YYYMGNLVMNN(1127.3184)YTPVQWRR | 1199.1656 | 4.0256 |
| decoy 10 | GYEWN(1592.6778)KSGMGEMHWYK | 1199.1833 | 3.8474 |
| decoy 11 | GFVMHSFTNSPCN(1620.7062)DSFK | 1199.1894 | 3.8474 |
| decoy 12 | FNDWALMN(1287.4151)QSVNMVWPTVR | 1199.1766 | 3.5206 |
| decoy 13 | FQFFN(1316.3959)DSRKNTHQFPHTK | 1199.1737 | 3.3116 |
| decoy 14 | LNASPQKN(1537.3879)HTYWVSTRR | 1199.1551 | 3.2378 |
| decoy 15 | YVDRTSYDYSRSHDN(1288.5189)WSR | 1199.1836 | 3.0834 |
| decoy 16 | PGRTCETDGHVTPLAN(1485.4418)QTR | 1199.1553 | 3.0834 |
| decoy 17 | N(1412.5759)YTAFESCMTNMLWKMR | 1199.1816 | 2.8765 |
| decoy 18 | DCYRYCPYKN(1562.6731)SSEYK | 1199.1802 | 0 |
| decoy 19 | VQDWWTN(1327.4822)ATMNVALNYWR | 1199.1874 | 0 |
| decoy 20 | DCWEYN(1395.4918)ATLKNNDKDWK | 1199.1609 | 0 |

**Figure 3.** For the input glycopeptide composition (labeled as target) DGGED*N*KTEEIFRPGGG- NMK + [Hex]3[HexNAc]4[Fuc]1, 20 decoy glycopeptide compositions were generated by GPE. Subsequently, GPE scored both the target and decoy glycopeptides against the ETD data, and they were ranked from high to low score as shown in this figure. The target glycopeptide, which is also the correct assignment, received the highest score of 61.7, outscoring all the other candidates. The scoring results of the other four incorrect glycopeptide candidates are summarized in Supporting Information, Table 1.

example shows how to use GPE: Correct and incorrect target glycopeptides can be readily differentiated by including a sufficient number of decoy glycopeptides in the scoring process, which are generated by GPE in an automated fashion.

**Is GPE Consistently Able to Identify the Correct Candidate, When It Is Present?** The above example illustrates that GPE can be used to effectively identify a correct target candidate among a large list of incorrect glycopeptides. To determine how consistently GPE could generate these kinds of successful results, we tested a larger data set. We employed GPE in analyzing a glycopeptide data set that contains ETD data of 77 distinct glycopeptides generated from multiple proteins (fetuin, IgG, HIV gp140, etc.). In these cases, all 77 spectra were manually assigned using the same procedure described above. After determining the correct assignment for each spectrum, four other (incorrect) "target" assignments were also generated. The software assigned 76 of the 77 MS/MS spectra to the correct glycopeptide compositions, demonstrating that the approach can consistently return the correct result, even when 20 decoys per candidate are scored. These results are expected when a high-quality algorithm is used for scoring glycopeptides, such as the one used in GPE, and the spectra are of high enough quality such that manual assignment is possible.

**Is GPE Effective at Identifying Misassigned Spectra?** We next tested whether GPE is capable of indicating that the incorrect target glycopeptides are incorrect when the true candidates are not present in the target list. The correct glycoprotein sequences that generated the ETD data were excluded from the search of target glycopeptide compositions, so that all the target glycopeptides were incorrect glycopeptide candidates from Titin. After the incorrect targets were input into GPE, they were scored along with 20 decoys per target. Only four out of the 77 ETD spectra were matched to the target glycopeptides that are incorrect, whereas 73 spectra were assigned to decoy glycopeptides. Therefore, the ratio of incorrect target matches to decoy matches, $N_{ic}/N_d$, is 0.055 (4/73) in this case. This value is very close to the target-to-decoy ratio of 0.05 (1/20).

**Comparison of the Predicted FDR to the True FDR.** Using the data above, we evaluated the true FDR for our data set of 77 spectra compared to the FDR that would be predicted by eq 4. When the correct glycopeptide compositions are included in the test, as mentioned above, 1 out of 77 assignments is a decoy match, and the FDR, according to eq 4, is predicted to be 1.36%. The actual FDR that is observed, on the other hand, is the number of incorrect assignments divided by the total assignments. In this case, only the decoy assignment is incorrect and the other 76 assignments are correct, so the observed FDR is 1.30% (1/77), which is closely approximated by the predicted FDR value. On the other hand, when the correct glycopeptide sequences are excluded from the target list, 73 of 77 assignments are decoy matches, which leads to a calculated FDR of 99.55%. (This calculation is done using eq 4: (73/77) × 1.05 = 0.995.) The actual FDR is 100% since all the assignments are incorrect. In both circumstances, the predicted FDRs are very close to the observed FDRs.

To further test if FDR values for small data sets can be accurately determined by our method, a proportion of the 77 ETD spectra were randomly selected, and for those spectra, the correct glycoprotein sequences were excluded for generating target glycopeptide candidates. For the remaining spectra, the correct glycoproteins were included in the generation of target compositions. Subsequently, GPE was employed to score each ETD spectrum against the corresponding target glycopeptides, and the number of decoy assignments was used to calculate FDR based on eq 4. The experiment was conducted at 12 different cases such that 0, 3, 5, 10, 20, 30, 40, 50, 60, 70, 73, 77, out of the 77 correct glycopeptide sequences were randomly excluded when their respective spectra were being scored. In this way, different numbers of incorrect assignments for the ETD data set were generated, and the predicted FDR using our method can be compared to the observed FDR at different levels. The comparison of the calculated versus observed FDRs for the 77 tested ETD spectra is illustrated in Figure 4A, where a correlation curve is made based on the blue data points. The least-squares fitting line has a slope that only deviates slightly from unity, and the curve has good linearity ($R^2$ above 0.99). These data demonstrate that for glycopeptide data set with a wide range of FDRs (ranging from 1.3%−100%), the FDR values can be determined accurately using GPE and the method that we developed.

In glycopeptide-based identifications, the MS/MS data set is frequently of a small size, and a robust method needs to be able to determine the FDRs for these types of data. To build a smaller glycopeptide data set, we randomly selected 35 ETD spectra from the entire data set, and performed the same experiment as described above, to test whether using our method, the FDRs at different levels can be measured with high accuracy for this limited size data set. The result is shown in Figure 4B where the correlation curve is fitted based on the blue data points; the best-fitting line between the predicted and observed FDRs has a slope that is close to 1 with $R^2$ still above


dx.doi.org/10.1021/ac502176n | *Anal. Chem.* 2014, 86, 9212−9219

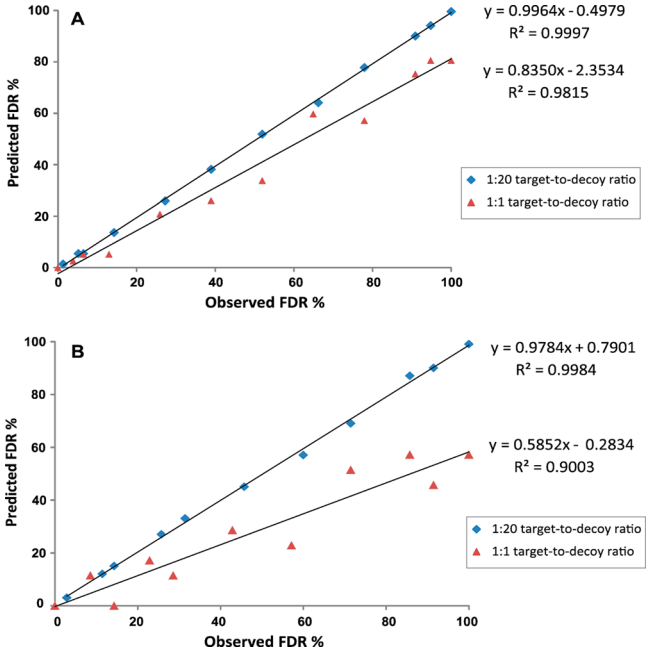



**Figure 4.** Lines that are fitted based on the blue data points: correlation curves between the predicted FDR values calculated using our method versus the observed FDR values that are manually verified. Lines that are fitted based on the red data points: correlation curves between the FDRs calculated by the common approach where an equal number of decoys are tested with the targets versus the observed FDRs that are verified manually. The FDRs are based on the analysis of ETD data sets of (A) 77 distinct glycopeptides and (B) 35 distinct glycopeptides using GPE program.

0.99. Therefore, these experiments prove that the developed method is accurate in measuring the FDRs in glycopeptide identifications, even for small glycopeptide data sets.

Finally, the accuracy of our method in predicting the FDRs was compared to that of the common approach where an equal number of decoy glycopeptides were tested with the target glycopeptides. For the same two data sets described above, correlation curves comparing the predicted versus observed FDRs, when using a 1:1 target-to-decoy ratio, are also shown in Figure 4. In this experiment, an equal number of decoy glycopeptides were generated by GPE based on the target candidates, and both the decoy and target glycopeptides are analyzed in the same way as described previously. These data sets are present in red. For the 77 tested ETD spectra, the $R^2$ of the curve is below 0.99, and the slope of the curve (0.83) deviates significantly from 1 (Figure 4A). Furthermore, using the conventional approach, the correlation between the predicted and observed FDRs becomes much worse when the size of the data set decreases, as evidenced by the correlation curve in Figure 4B that has a $R^2$ of only 0.90 and a flat slope of 0.58. The slope of the curves reflect the ratio of predicted FDRs to true FDRs, and the values, which are significantly less than 1, indicate that the number of false positive assignments would be considerably underestimated using the conventional approach. By contrast, the FDRs are predicted accurately using our method, especially under circumstances where only a small glycopeptide data set is available.

## CONCLUSION

False discovery rate (FDR) is an important measurement of the confidence of glycopeptide assignments when MS/MS data of glycopeptides are analyzed. In order to accurately determine the FDR of glycopeptide identifications, we developed a software program, GlycoPep Evaluator, to generate abundant decoy glycopeptide compositions and to score the target and decoy glycopeptide candidates in measuring the FDR. The target-to-decoy ratio is 1:20 so that, even for a small number of target glycopeptide sequences, sufficient decoy glycopeptides are available for scoring; hence, false-positive identifications can be better contained. Moreover, FDRs can be measured with high accuracy using GPE for small data sets, which are commonly seen in glycoproteomics where tens to hundreds of spectra are scored, as opposed to thousands of spectra scored in a proteomics experiment. The functionality of GPE in generation of decoy glycopeptide candidates can be combined with any other data analysis tools that score ETD data of glycopeptides, so that FDRs can be accurately determined.

## ASSOCIATED CONTENT

### Supporting Information

Table of the scores of the remaining 4 targets, their 80 decoys. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*Phone: 785-864-3015. Fax: 785-864-5396. E-mail: hdesaire@ku.edu.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors acknowledge funding from the National Institutes of Health (Grants R01GM103547 and R01AI094797) to H.D., and an American Chemical Society Division of Analytical Chemistry Graduate Fellowship, sponsored by Eastman Chemical Company, to Z.Z.

## REFERENCES

(1) Apweiler, R.; Hermjakob, H.; Sharon, N. *Biochim. Biophys. Acta-Gen. Subj.* **1999**, *1473*, 4−8.
(2) Khoury, G. A.; Baliban, R. C.; Floudas, C. A. *Sci. Rep.* **2011**, *1*.
(3) Banerjee, S.; Vishwanath, P.; Cui, J.; Kelleher, D. J.; Gilmore, R.; Robbins, P. W.; Samuelson, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11676−11681.
(4) Banks, D. D. *J. Mol. Biol.* **2011**, *412*, 536−550.
(5) Furukawa, K.; Ohkawa, Y.; Yamauchi, Y.; Hamamura, K.; Ohmi, Y. *J. Biochem.* **2012**, *151*, 573−578.
(6) Van den Steen, P.; Rudd, P. M.; Dwek, R. A.; Opdenakker, G. *Crit. Rev. Biochem. Mol. Biol.* **1998**, *33*, 151−208.
(7) Coelho, V.; Krysov, S.; Ghaemmaghami, A. M.; Emara, M.; Potter, K. N.; Johnson, P.; Packham, G.; Martinez-Pomares, L.; Stevenson, F. K. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 18587−18592.
(8) Wang, Y.; Tan, J.; Sutton-Smith, M.; Ditto, D.; Panico, M.; Campbell, R. M.; Varki, N. M.; Long, J. M.; Jaeken, J.; Levinson, S. R.; Wynshaw-Boris, A.; Morris, H. R.; Le, D.; Dell, A.; Schachter, H.; Marth, J. D. *Glycobiology* **2001**, *11*, 1051−1070.
(9) Li, Y.; Tian, Y. A.; Rezai, T.; Prakash, A.; Lopez, M. F.; Chan, D. W.; Zhang, H. *Anal. Chem.* **2011**, *83*, 240−245.
(10) Leymarie, N.; Zaia, J. *Anal. Chem.* **2012**, *84*, 3040−3048.
(11) Desaire, H. *Mol. Cell. Proteomics* **2013**, *12*, 893−901.

(12) Hong, Q.; Lebrilla, C. B.; Miyamoto, S.; Ruhaak, L. R. *Anal. Chem.* **2013**, *85*, 8585−93.
(13) Desaire, H.; Hua, D. *Int. J. Mass Spectrom.* **2009**, *287*, 21−26.
(14) Woodin, C. L.; Maxon, M.; Desaire, H. *Analyst* **2013**, *138*, 2793−2803.
(15) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. *J. Proteome Res.* **2008**, *7*, 1650−1659.
(16) Premier Biosoft. http://www.premierbiosoft.com/glycan/index.html (accessed April 30, 2014).
(17) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H. R.; Dell, A. *J. Proteomics Res.* **2007**, *6*, 3995−4005.
(18) Apte, A.; Meitei, N. S. *Methods Mol. Biol.* **2010**, *600*, 269−81.
(19) Chandler, K. B.; Pompach, P.; Goldman, R.; Edwards, N. *J. Proteome Res.* **2013**, *12*, 3652−3666.
(20) Bruker Daltonics. http://glycomics.ccrc.uga.edu/GlycomicsPortal/showEntry.action?id=116 (accessed April 30, 2014).
(21) Strum, J. S.; Nwosu, C. C.; Hua, S.; Kronewitter, S. R.; Seipert, R. R.; Bachelor, R. J.; An, H. J.; Lebrilla, C. B. *Anal. Chem.* **2013**, *85*, 5666−5675.
(22) Mayampurath, A.; Yu, C. Y.; Song, E. W.; Balan, J.; Mechref, Y.; Tang, H. X. *Anal. Chem.* **2014**, *86*, 453−463.
(23) Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. *Anal. Chem.* **2012**, *84*, 4821−4829.
(24) Zhu, Z.; Hua, D.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.* **2013**, *85*, 5023−5032.
(25) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. *Nat. Methods* **2007**, *4*, 787−797.
(26) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207−214.
(27) Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 47−50.
(28) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2008**, *7*, 29−34.
(29) Nesvizhskii, A. I. *J. Proteomics* **2010**, *73*, 2092−2123.
(30) Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *J. Proteome Res.* **2006**, *6*, 392−398.
(31) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. *Mol. Cell. Proteomics* **2009**, *8*, 2405−2417.
(32) Reidegeld, K. A.; Eisenacher, M.; Kohl, M.; Chamrad, D.; Korting, G.; Blueggel, M.; Meyer, H. E.; Stephan, C. *Proteomics* **2008**, *8*, 1129−1137.
(33) Wang, G.; Wu, W. W.; Zhang, Z.; Masilamani, S.; Shen, R. F. *Anal. Chem.* **2009**, *81*, 146−159.
(34) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111−1120.
(35) Chalkley, R. J. *J. Proteome Res.* **2013**, *12*, 1062−1064.
(36) Go, E. P.; Chang, Q.; Liao, H. X.; Sutherland, L. L.; Alam, S. M.; Haynes, B. F.; Desaire, H. *J. Proteome Res.* **2009**, *8*, 4231−4242.
(37) Zhu, Z.; Su, X.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.* **2013**, *85*, 8403−8411.
(38) Zhu, Z.; Go, E. P.; Desaire, H. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1012−1017.
(39) Alley, W. R.; Mechref, Y.; Novotny, M. V. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 161−170.
(40) Wada, Y.; Azadi, P.; Costello, C. E.; Dell, A.; Dwek, R. A.; Geyer, H.; Geyer, R.; Kakehi, K.; Karlsson, N. G.; Kato, K.; Kawasaki, N.; Khoo, K. H.; Kim, S.; Kondo, A.; Lattova, E.; Mechref, Y.; Miyoshi, E.; Nakamura, K.; Narimatsu, H.; Novotny, M. V.; Packer, N. H.; Perreault, H.; Peter-Katalinic, J.; Pohlentz, G.; Reinhold, V. N.; Rudd, P. M.; Suzuki, A.; Taniguchi, N. *Glycobiology* **2007**, *17*, 411−422.
(41) Zhang, Y.; Go, E. P.; Desaire, H. *Anal. Chem.* **2008**, *80*, 3144−3158.
(42) Cooper, C. A.; Gasteiger, E.; Packer, N. H. *Proteomics* **2001**, *1*, 340−349.