Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

When can glycopeptides be assigned based solely on high-resolution mass spectrometry data?

Heather Desaire*, David Hua

Department of Chemistry, University of Kansas, 2030 Becker Drive, Lawrence, KS 66047, United States

ARTICLE INFO

Article history:
Received 2 July 2008
Received in revised form
18 November 2008
Accepted 1 December 2008
Available online 9 December 2008

Keywords:
Glycoproteomics
Glycoprotein
Glycan
High-resolution
GlycoPep DB

ABSTRACT

Glycoproteomics is an emerging science that shows promise in applications such as biomarker discovery and biopharmaceutical development. One central technique in glycoproteomic analysis is analyzing glycopeptides by mass spectrometry. This challenging technique is still under development, and methods to simplify the data analysis are greatly needed. One potentially attractive analysis approach would be to assign a significant portion of the glycopeptide compositions using high-resolution MS data. In the work described herein, we ask the question: Under what circumstances is it possible to assign glycopeptides to MS data, using only high-resolution mass spectra? Variables investigated include the number of glycosylation sites on the protein, the potential diversity of the glycans attached to the protein, and the mass accuracy obtained. This work outlines guidelines for when it is (and is not) appropriate to rely heavily on high-resolution mass measurements to assign glycopeptide compositions; such guidelines are potentially useful for anyone conducting glycopeptide analysis by mass spectrometry.

1. Introduction

Glycosylation is known to significantly impact the structure and function of proteins, including protein folding [1], receptor binding [2], and metabolic clearance [3,4]. While the glycosylation of numerous proteins has been studied, much more research needs to be done in this important area. Glycosylation analysis of proteins can be accomplished by deglycosylating the protein of interest and analyzing released glycans [5–7], or by digesting the protein(s) of interest with proteases and analyzing glycopeptides [8,9]. The latter approach is technically more challenging, but it provides key information about the attachment sites of each of the analyzed glycans. This information is often critical in biomarker discovery [10], vaccine design [11], and structure/function studies of glycoproteins [12–16].

Glycopeptides are typically analyzed by mass spectrometry, and a key challenge in this research area is correctly assigning the glycopeptide compositions to the MS data. The fundamental challenge is that two unknowns, the peptide mass and the glycan mass, must be correctly identified [8]. A typical strategy for accomplishing this task is to first assign one of the two unknown portions, using MS/MS data, and then to use high-resolution MS data to complete the mass assignment. For example, the Peptoonist program first determines the peptide assignment by searching for the [Peptide + HexNAc + H]$^+$ ion, along with b and y ions of the peptide, in the MS/MS data of a glycopeptide, and uses these ions to identify the peptide portion of the glycopeptide; then it utilizes high-resolution MS data along with the peptide mass to identify the mass of the glycan portion [17]. This strategy has been implemented previously by several other groups [18–20]. The key limitation to this approach is that MS/MS data are required. This requirement is particularly problematic in glycopeptide-based analyses because glycosylated peptides are known to be present in low abundance in the MS data, compared to non-glycosylated peptides [19,21]. Therefore, many glycopeptide peaks will be skipped in data-dependent LC-MS/MS runs unless multiple different runs are conducted with complementary data-dependent acquisition strategies; MALDI-MS/MS is also not feasible on very low-abundant species.

If high-resolution mass spectral data could be used to correctly assign the full glycopeptide composition, without relying on MS/MS data, glycopeptide analysis could be expedited significantly, and more low-abundance peaks, which do not produce quality MS/MS data, could be assigned. Of course considerable anecdotal evidence is available in the literature to demonstrate that relying solely on high-resolution data is problematic, because sometimes two different [glycan + peptide] combinations can add to nearly identical masses [17,20]. Certainly, cases exist in which some glycopeptides cannot be unequivocally assigned by high-resolution mass alone, but the question to be answered here is: When (if ever) could high-resolution data be sufficient in accurately assigning glycopeptide masses? We answer this question by producing lists of all possible glycopeptide masses for two model proteins and determining how

* Corresponding author. Tel.: +1 785 864 3015.
E-mail address: hdesaire@ku.edu (H. Desaire).

1387-3806/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2008.12.001

many of these masses are unique, within a given mass tolerance. We specifically investigated the effects of the number of glycosylation sites on the protein, and the effects of the mass accuracy of the mass spectrometer. Finally, we experimentally determined the mass accuracy of an LTQ-FT-MS and a MALDI-TOF/TOF MS in analyzing a complex glycopeptide sample. Using these experimentally determined values, it was possible to approximate the percent of false positives one would obtain if glycopeptide compositions were assigned using high-resolution MS data alone, not in conjunction with MS/MS data.

2. Experimental

The sequence of human α-1-acid glycoprotein (AGP) (accession number P02763) was obtained from the SwissProt database, at http://www.expasy.ch/sprot/. The sequence of the protein CON-S gp140ΔCFI (CON-S) was obtained from reference [22]. Both protein sequences are shown in Fig. 1. Glycosylation sites on these proteins were identified by searching for the consensus sequence N-X-T/S, where X is any amino acid except proline. After the glycosylation sites were identified, a list of tryptic peptides containing those glycosylation sites was generated. The list contained all possible tryptic peptides that had zero or one missed cleavages that included at least one glycosylation site. For each protein, the neutral, monoisotopic masses of all the tryptic peptides were calculated to four decimal places. An array of the [peptide masses + compositions] was used as input data when constructing the database of all possible glycopeptide masses.

Input data for glycans were also acquired. A total of 187 glycan compositions and their corresponding masses were exported from the database GlycoPep DB [23]. These glycans are all N-linked glycans, and each has been previously described in the literature as being present on mammalian glycoproteins. Each glycan composition's monoisotopic mass was also calculated to four decimal places.

Two different databases of glycopeptide masses were generated separately for the two glycoproteins, AGP and CON-S. For each database, the same list of glycans was used, and the tryptic peptides (described above) for the corresponding protein were used. All the [glycan + peptide] combinations were included in a list of all possible glycopeptides. This was accomplished by coding a series of short functions in the open-source database software PostgreSQL (version 8.3.2), to create tables adding each peptide in the peptide array to each glycan in the glycan array. The data were output to an Excel file for further processing.

After the list of all possible glycopeptides and their corresponding masses was generated, the data were processed by grouping the glycopeptides into several categories. To achieve this objective, the list of glycopeptides and their masses was sorted from lowest to highest mass. For each glycopeptide, the mass difference between the glycopeptide of interest and the next larger glycopeptide was calculated. Additionally, the mass difference between each glycopeptide and the next smaller glycopeptide was also calculated. The smaller of these two mass differences was used in each case, to group the data, into bins of 1 ppm, 5 ppm, 10 ppm, 20 ppm, 100 ppm and 150 ppm mass windows. For example, a glycopeptide whose mass was within 1 ppm of another glycopeptide mass, was included in the 1 ppm mass window. This data were used to calculate the percent of uniquely identifiable glycopeptides for each mass window.

3. Results and discussion

Can high-resolution MS (HR-MS) be used to unequivocally identify glycopeptide compositions for a given glycoprotein? If so, what level of mass accuracy would be needed in order to produce high-confidence assignments? To address these questions, we chose two model proteins to analyze, α-1-acid glycoprotein (AGP) and the 2001 Group M Consensus Sequence for the HIV-1 Envelop protein (CON-S). The sequences for the two proteins, along with their glycosylation sites, are shown in Fig. 1.

These two analytes were chosen to represent two important classes of glycoproteins. CON-S is a very heavily glycosylated protein, with 31 potential glycosylation sites spread over 19 tryptic peptides [11,24]. The glycosylation on this protein likely contributes to its enhanced efficacy as an HIV-1 vaccine candidate, compared to other similar proteins [11]. Furthermore, studying the glycosylation on this and similar proteins is a current focus for vaccine researchers. Outside the HIV field, glycopeptide analysis for a protein such as this represents a very significant challenge, due to the large number of glycosylation sites. By contrast, α-1-acid glycoprotein (AGP) is a much less challenging analyte, with only five glycosylation sites. It was chosen to represent a less complex glycoprotein, where there are only five possible peptide sequences that would account for any resulting glycopeptides.

Using these two proteins to represent a "complex" case and a "standard" case, we sought to determine whether high-resolution MS analysis alone would be sufficient to unequivocally identify glycopeptide compositions from these proteins. Glycopeptide analysis involves identifying both the peptide composition and the glycan composition for the glycopeptide. So in essence, there are two unknowns (the peptide mass and the glycan mass) that need to be identified. If no information was available about the protein or the glycans, it would be impossible to infer what portion of the mass was due to the protein, and what portion of the mass was due to the glycan. However, when the protein sequence of the analyte is known, the possible combinations of [peptide mass + glycan mass] are now reduced, since there are a limited number of possible peptide masses. These masses can be calculated by conducting a theoretical digest on the protein.

One method of determining the likelihood that high-resolution mass data alone would be sufficient to uniquely identify a glycopeptide composition is to calculate the masses of all the possible glycopeptides and then determine how many of these species are uniquely identified by their mass. This is the approach we used, and it is described in Fig. 2. The first step is to determine what should be counted in the list of all possible peptides. This is fairly straight forward. One can calculate the mass of the tryptic peptides that contain glycosylation sites. In this case, the only additional factors to consider are whether to include missed tryptic cleavages and whether to also consider peptide modifications, such as deamidation, oxidation, etc. Since glycans are known to provide a steric barrier that can prevent proteolysis from occurring, we assumed that each of the proteins could potentially undergo up to one missed cleavage. We did not include any possible peptide modifications, such as deamidation or formic acid adduction, in the peptide lists because while modifications do occur, their incidence can be minimized by careful control of sample preparation conditions. (Oxidation, however, commonly occurs as a natural protein modification, and the glycosylation on an oxidized glycopeptides could be easily mis-assigned, since hexose and fucose differ only by one oxygen. This issue is not resolved by high-resolution MS data.)

For AGP, the five glycosylation sites are located on five unique tryptic peptides. However, some of these tryptic peptides are adjacent to each other, so only seven additional peptides are added to the data set, when all glycopeptides with one missed cleavage are accounted for. For CON-S, which contains 31 glycosylation sites on 19 tryptic peptides, the total number of possible peptide masses to be considered is 47. This number includes the 19 glycopeptides containing zero missed cleavages and an additional 28 peptides that each contain at least one glycosylation site and one missed cleavage.

(a) Con-S protein sequence

MRVRGIQRNCQHLWRWGTLI LGMLMICSAAENLWVTVYYG VPVWKEA **N**TTLFCASDAKAY
DTEVHNVWATHACVPTDPNP QEIVLE **N**VTENFNMWKNNMV EQMHEDIISLWDQSLKPCVK
LTPLCVTL**N**CTNV**N**VT**N**TT**N** NTEEKGEIK**N**CSF**N**ITTEIR DKKQKVYALFYRLDVVPIDD
NN**NN**SSNYRLINC**N**TSAITQ ACPKVSFEPIPIHYCAPAGF AILKCNDKKF **N**GTGPCK**N**VS
TVQCTHGIKPVVSTQLLL**N**G SLAEEEIIIRSE**N**ITNNAKT IIVQL**N**ESVEI**N**CTRPN**N**NT
RKSIRIGPGQAFYATGDIIG DIRQAHC **N**ISGTKW**N**KTLQQ VAKKLREHFN**N**KTIIFKPSS
GGDLEITTHSFNCRGEFFYC **N**TSGLF**N**STWIG**N**GTKNN**N**N T**N**DTITLPCRIKQIINMWQG
VGQAMYAPPIEGKITCKS**N**I TGLLLTRDGGN**N**NT**N**ETEIF RPGGGDMRDNWRSELYKYKV
VKIEPLGVAPTKAKLTVQAR QLLSGIVQQQSNLLRAIEAQ QHLLQLTVWGIKQLQARVLA
VERYLKDQQLEIWD**N**MTWME WEREIN**N**YTDIIYSLIEESQ NQQE

(b) AGP protein sequence

MALSWVLTVL SLLPLLEAQI PLCANLVPVP IT **N**ATLDQIT GKWFYIASAF
RNEEY**N**KSVQ EIQATFFYFT P**N**KTEDTIFL REYQTRQDQC IY**N**TTYLNVQ
RE**N**GTISRYV GGQEHFAHLL ILRDTKTYML AFDVNDEKNW GLSVYADKPE
TTKEQLGEFY EALDCLRIPK SDVVYTDWKK DKCEPLEKQH EKERKQEEGE S

Fig. 1. Protein sequences used for this study. (a) CON-S. (b) AGP. Glycosylation sites are in bold.

In calculating all possible glycopeptide masses, obtaining the list of possible peptides is only half the challenge. One must also procure a list of glycan masses. This is Step 2 in Fig. 2. Theoretically, a list of glycan masses could contain a virtually limitless number of species. For example, when searching for glycan compositions in GlycoMod [25], a well-known tool for generating glycan structures and masses, 491 different glycan compositions are reported between the masses of 2000 and 2100. The total number of glycan compositions generated by GlycoMod is in the thousands. Arguably, this is not the best set of glycans to use in the present analysis, since many of the glycan compositions are generated by GlycoMod are not likely to be biologically relevant. That is, they have not been described in the literature, and they cannot be logically constructed based on the enzymatic glycosylation and deglycosylation processes that occur in the Golgi. Instead of using a large and potentially meaningless list of glycans, we chose to use the list of glycans in the GlycoPep DB database. GlycoPep DB contains an actively curated database of glycans, all of which have been described in the literature [23]. This database has been used to fully characterize several glycoproteins [8,11,21,24], and its entries include all of the typical types of N-linked glycans: high mannose, complex, hybrid, fucosylated and non-fucosylated, sialylated and non-sialylated, phosphorylated, sulfated, etc. Currently, there are 187 entries with unique masses in GlycoPep DB, and the entire list of glycans was selected for use in this project. (After the analyses is completed using the list of glycans from GlycoPep DB, we discuss how the results of the analysis would vary if a more extensive list of glycans is used.)

Once the lists of possible peptides and possible glycans were obtained, a series of functions was written in PostgreSQL to construct the list of all possible glycopeptide masses, by adding every entry in the "peptide inputs" list to every entry in the "glycan inputs" list, as shown in Step 3 of Fig. 2. For example, the protein CON-S, had 47 peptide inputs, and 187 glycan inputs, so the total number of possible glycopeptide masses for this protein is 47 × 187, or 8789. This list represents all the possible glycopeptide masses for CON-S, with the restrictions that the peptides contain 0 or 1 missed cleavages and all the glycans are accounted for in the GlycoPep DB database. A portion of the 8789 entries are shown in Table 1. These data are sorted by mass, and the mass difference between each entry is calculated.

Once all the possible glycopeptide masses for CON-S were calculated, we assessed whether the high-resolution MS data alone would be sufficient to unequivocally identify glycopeptides from this protein by first subdividing each of the glycopeptides into groups based on their proximity to another glycopeptide. The proximities of interest were 1 ppm, 5 ppm, 10 ppm, 20 ppm, 100 ppm, 150 ppm, and those whose masses were more than 150 ppm from the next-nearest glycopeptide mass. The percent that fell into each category is shown in Fig. 3. For CON-S, 29% of the glycopeptides in the database were within 10 ppm of another glycopeptide. Therefore, if one could only measure glycopeptide masses with an accuracy of ±5 ppm (which creates a 10 ppm mass window), then 29% would not be uniquely identified. In other words, only 71% of the glycopeptides in the CON-S database could be uniquely identified by their mass alone; worse yet, it would be impossible to tell *which* glycopeptides could be uniquely identified in advance, without constructing a data table such as the one described in Fig. 2.

The data in Fig. 3 demonstrate that for a very complex glycoprotein like CON-S, high-resolution MS analysis alone would not suffice to uniquely identify a high percentage of glycopeptides, unless the mass accuracy is quite high. For example, if the mass accuracies for all the assigned peaks were within ±0.5 ppm (making a mass window of 1 ppm), then 98% of the glycopeptides could be uniquely

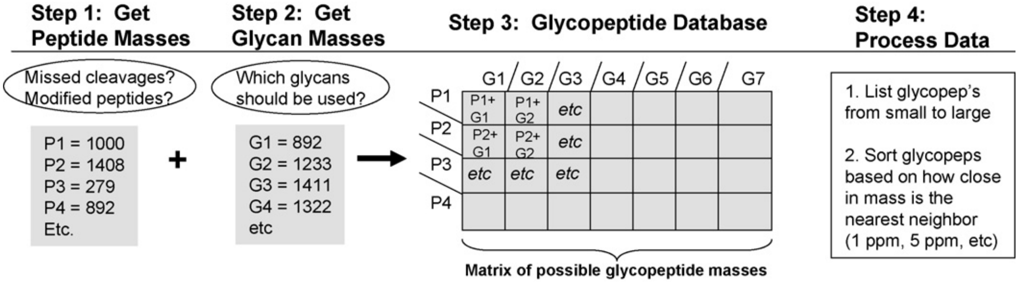
Step 1: Get Peptide Masses
Step 2: Get Glycan Masses
Step 3: Glycopeptide Database
Step 4: Process Data
Missed cleavages? Modified peptides?
Which glycans should be used?
P1 = 1000
P2 = 1408
P3 = 279
P4 = 892
Etc.
+
G1 = 892
G2 = 1233
G3 = 1411
G4 = 1322
etc
G1
G2
G3
G4
G5
G6
G7
P1
P2
P3
P4
P1+ G1
P1+ G2
P2+ G1
P2+ G2
etc
Matrix of possible glycopeptide masses
1. List glycopep's from small to large
2. Sort glycopeps based on how close in mass is the nearest neighbor (1 ppm, 5 ppm, etc)

Fig. 2. Overall strategy for determining how many glycopeptides can be uniquely identified by their mass.

Table 1
Sample output data combining peptides and glycans.

Peptide composition	Glycan composition	Mass of glycopep	♦ Mass (ppm) vs. next smaller	♦ Mass (ppm) vs. next larger
GEIKNCSFNITTEIR	[Hex]6[HexNAc]6[Fuc]1[NeuNAc]2	4699.9198	219	7
EINNYTDIIYSLIEESQNQQE	[Hex]4[HexNAc]6[Fuc]1[NeuNAc]1	4699.9542	7	20
LTPLCVTLNCTNVNVTNTTNNTEEK	[Hex]3[HexNAc]6[Fuc]1	4700.0462	20	14
TIIVQLNESVEINCTRPNNNTRK	[Hex]6[HexNAc]5	4700.1115	14	147
NNNNTNDTITLPCR	[Hex]6[HexNAc]5[Fuc]1[NeuGc]2	4700.8038	147	9
DQQLEIWDNMTWMEWER	[Hex]4[HexNAc]5[Fuc]1[NeuNAc]2	4700.8451	9	18
IGPGQAFYATGDIIGDIRQAHCNISGTK	[Hex]6[HexNAc]3[SO3]2	4700.9298	18	3
YLKDQQLEIWDNMTWMEWER	[Hex]6[HexNAc]5	4700.9444	3	24
TIIVQLNESVEINCTRPNNNTR	[Hex]5[HexNAc]5[NeuNAc]1	4701.0591	24	418
IGPGQAFYATGDIIGDIRQAHCNISGTK	[Hex]4[HexNAc]5[SO3]1	4703.0262	418	9
IGPGQAFYATGDIIGDIRQAHCNISGTK	[Hex]7[HexNAc]3	4703.069	9	174
GEFFYCNTSGLFNSTWIGNGTK	[Hex]5[HexNAc]4[NeuNAc]2	4703.8891	174	256

Note: For the peptides that contain multiple potential glycosylation sites, no information is obtainable from the high-resolution mass data about which of the site(s) are occupied or how the glycans are distributed among the occupied sites.

identified. While it is certainly possible to obtain mass errors of less than 1 ppm on an FT-MS instrument, typical mass errors for glycopeptides analyzed by FT-MS in the literature are higher, as demonstrated below.

Does the situation improve if a less complex protein is analyzed? To answer this question, a similar analysis was completed for AGP. The results of the analysis are shown in Fig. 3b. In this case, about 91% of the glycopeptides had unique masses, within a 10 ppm mass window. Therefore, if MS peaks for glycopeptides were generated at a mass accuracy of ±5 ppm, there would be a reasonably high probability that the masses for these glycopeptides could be uniquely identified, based solely on their mass. However, if the mass error of the glycopeptides drifts much above a 10 ppm mass window (±5 ppm), the confidence in the mass assignment goes down quickly. For an analysis that generates data with up to ±10 ppm mass error (a 20 ppm mass window), only 79% of the glycopeptides could be uniquely identified.

In addition to the fact that some of the species would not be uniquely identifiable as a single glycopeptide, one additional potential problem is that using only high-resolution data, there is no guarantee that the mass spectral peaks, which are correlated to glycopeptide compositions, originate from a glycopeptide. Any MS peak that matches a glycopeptide mass could also originate from a non-glycosylated peptide from the protein of interest or a peptide or glycopeptide from another contaminating species. One must consider the possibility of these contaminants being present.

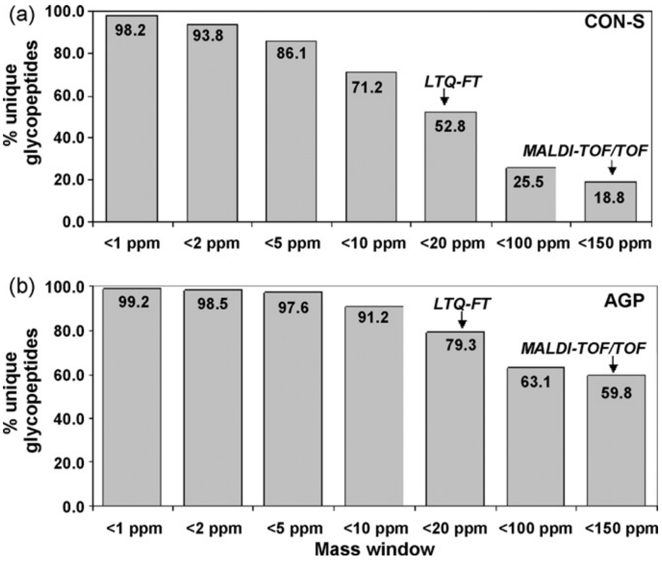

Fig. 3. Data showing what percent of the glycopeptides can be uniquely identified by their mass, as the mass window is varied. (a) Results for CON-S. (b) Results for AGP. Both graphs also show the experimentally determined mass accuracy for an LTQ-FTMS and a MALDI-TOF/TOF MS, as described in Fig. 4.

In summary, if high mass accuracies (±5 ppm) are obtainable on glycopeptide ions, glycopeptide composition data could be uniquely assigned for moderately complex glycoproteins, with a moderately low false-positive rate, about 9% for AGP. (This ignores the fact that the MS peaks are not verified to be glycopeptides.) Many instrument vendors advertise that their instruments can obtain better than a 5 ppm mass accuracy, but these measurements are typically made on pure standards, not complex mixtures of glycopeptides. When the goal is to analyze all the heterogeneity present in a complex mixture, more ions are typically injected into the mass analyzer, so the low-abundant species can be detected. These low-abundant species cannot be measured as accurately, especially when many other species are simultaneously present at higher abundance.

To determine whether it is realistic to measure most glycopeptides with ±5 ppm mass accuracy, we experimentally determined the average mass error for CON-S glycopeptides, which were analyzed by both a MALDI-TOF/TOF and an LTQ-FT-MS. These data have recently been published [11], and they provide a useful data set for determining realistic mass errors of glycopeptide ions. In the data set of all the glycopeptide ions analyzed (provided as supplemental information in reference [11]), there are 82 CON-S glycopeptide ions that were identified by both the MALDI-TOF/TOF and LTQ-FT-MS instruments. These 82 ions were chosen as a useful experimental data set for three reasons: First, these species were detected on both instruments, so they allow for a head-to-head comparison between the two instruments. Second, they represent the highest confidence of the mass assignments in the data set, because the two instruments that were used to identify them provide complementary information in their MS/MS data [24]. Finally, this data set represents a reasonably sized group of glycopeptide species, as detected during an analysis of a relevant glycopeptide sample.

A full list of each of the glycopeptide ions, along with their mass errors, is provided as supplementary data. Fig. 4 shows a scatter plot of the mass errors for each of the 82 ions, along with bars that show the median mass error and the mass accuracy needed to account for 98% of the ions. For the FT-MS data (Fig. 4a), the median mass error was 4.1 ppm, which is much better than that of the TOF/TOF, which had a median mass error of 13 ppm for the same set of ions; see Fig. 4b. In addition, 30% of the peaks from the FT-MS data had mass errors of greater than ±5 ppm; while 88% of the peaks in the MALDI data had more than 5 ppm mass error. (Data not shown.) If the mass errors of this analysis are representative of typical LC-FT-MS and MALDI-TOF/TOF analyses of complex glycopeptides mixtures, one must expect that a significant fraction of the mass errors for glycopeptide analytes are outside a ±5 ppm mass window. Therefore,

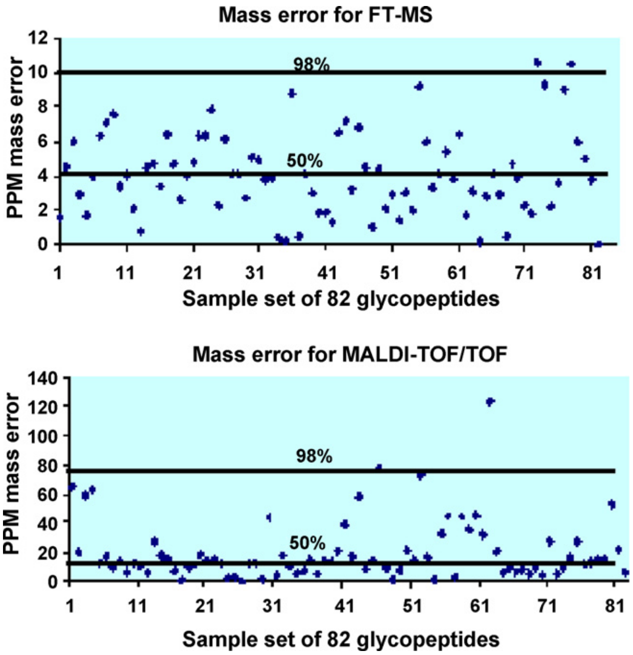

Fig. 4. Experimentally determined mass error for 82 previously assigned glycopeptides (from reference [11]) on the (a) LTQ-FTMS and (b) MALDI-TOF/TOF MS. The median mass error (50%) and the mass error that encompasses 98% of the assigned glycopeptides are shown on the graphs.

even for a simple glycoprotein like AGP, many of the glycopeptide ions would not be unambiguously assignable, since the number of uniquely assignable ions drops off significantly when the mass error is larger than ±5 ppm.

The FT-MS clearly produced better mass accuracy for the test glycopeptides described above. One would expect that this would lead to significantly more unambiguous assignments compared to data obtained on the MALDI TOF/TOF. To test this hypothesis, we determined what level of mass accuracy for each instrument was sufficient to include 98% of the assigned peaks and then we assessed what percent of the glycopeptides' masses in the theoretical list of glycopeptides is uniquely assigned for those mass accuracies. For the FT-MS, a mass error of ±10 ppm (a 20 ppm mass window) encompassed 98% of the mass errors for the assigned glycopeptides in the data set; see Fig. 4a. For the MALDI-TOF/TOF, 98% of the assignments were within ±75 ppm mass error (a 150 ppm window), as shown in Fig. 4b. The data in Fig. 3 indicate the number of unique assignments for both a 20 ppm mass window and a 150 ppm mass window. While it would seem that the 150 ppm data would include many more ions with more than one reasonable assignment, compared to data with 20 ppm mass accuracy, the data demonstrate otherwise. Only about 1/3 of the total ions become uniquely assignable in going from the 150 ppm mass window to the 20 ppm mass window, for CON-S. For the smaller protein, only 20% of the total ions become uniquely assignable based on their mass, in going from the 150 ppm window to the 20 ppm window. Therefore, in terms of identifying glycopeptides based on unique masses, the data in Fig. 3a and b indicate that only a moderate improvement is made in going from ±75 ppm mass error to ±10 ppm. By contrast, a substantial improvement is made in going from ±10 ppm to ±1 ppm mass error. For AGP, 98.5% of the glycopeptides could be uniquely identified, at ±1 ppm mass error. Even for a complex glycoprotein like CON-S, a substantial number of glycopeptides with unique mass assignments (94%) are present at the ±1 ppm level (a 2 ppm mass window).

If high-resolution MS is to be used to unequivocally identify glycopeptide by mass alone, every effort should be taken to assure that the mass error is minimized. The best way to determine the mass error necessary to produce a high level of unambiguous assignments would be to calculate all possible peptide and glycan masses, for the given protein, as described herein, and experimentally determine what mass error was necessary for a high likelihood of uniquely identifying a glycopeptide, based on the glycopeptide's mass. A less-well-tested but simpler approach would be to use the data herein as a guide. For less complex glycoproteins, (up to 5 glycosylation sites), a mass error of ±2.5 ppm (a 5 ppm mass window) would be good enough to provide a reasonable likelihood of uniquely identifying the glycopeptides, assuming that the number of peptide modifications was minimized; the protein underwent at most, one missed cleavage; and the glycans assigned were normal N-linked mammalian glycans such as those found in GlycoPep DB. For more complex proteins, such as CON-S, or other similar proteins with 20 or more glycosylation sites, one can safely assume that it is unlikely that even the best high-resolution mass data would be sufficient to accurately assign the glycopeptide masses, based on the calculations presented herein. The presence of anomalous peptide cleavages can further complicate the issue.

In terms of assessing whether or not the high-resolution data are good enough to assign glycopeptides without further confirmation, the most critical assumption is that the glycans assigned to the glycopeptide ions within the range of the diversity of glycans used in this analysis (187 unique species.) If the diversity of glycans potentially present in the sample is large – that is, if it is possible that any of the thousands of glycan compositions retrievable in GlycoMod are realistic assignments for the glycan mass – then the guidelines described herein would severely underestimate the mass accuracy needed to uniquely identify the glycopeptides, based on high-resolution mass data alone. In other words, if the potential variability in the glycan composition is as large as the data set found in searching GlycoMod, then even simple proteins, with just a few glycosylation sites, would have numerous feasible mass assignments for any detectable glycopeptide ion. Therefore, researchers should not rely solely on high-resolution MS data to assign both peptide and glycan compositions when using a large, unrestricted set of potential glycans to assign the glycopeptide.

4. Conclusion

If one were to consider every possible glycan structure generated via GlycoMod, HR-MS could not be used to uniquely assign glycopeptide compositions, based only on the high-resolution mass. In cases where it is appropriate to limit the glycan search to include a restricted set of glycans, the percent of unambiguous glycopeptide identifications from HR-MS varies with the complexity of the protein. For complex glycoproteins, such as CON-S, 2% of the possible glycopeptide assignments would have ambiguous assignments, even when data are acquired at sub-ppm mass accuracy. Rather simple proteins, such as AGP, require ±2.5 ppm mass accuracy (a 5 ppm mass window) for HR-MS alone to produce unambiguous assignments a high percentage (∼98%) of the time. Both the MALDI-TOF/TOF and the ESI-LTQ-FTMS can provide less than 2.5 ppm mass error. However, for typical HR-MS data of glycopeptide mixtures acquired on the LTQ-FT (up to 10 ppm mass error) or a MALDI-TOF/TOF (up to 75 ppm mass error), neither the AGP nor CON-S glycoproteins could be assigned with high confidence, using the high-resolution MS data alone. Surprisingly, not much value is added by increasing the mass accuracy from 75 ppm (the level for the MALDI-TOF/TOF to 10 ppm (provided by the LTQ-FT). For both proteins, many of the ambiguous glycopeptide assignments that were present in the 75 ppm data were also present in the 10 ppm data.

In summary, this study demonstrates that only a few specialized cases exist where high mass accuracy alone can be relied upon to uniquely identify glycopeptides. If a simple glycoprotein that is not known to undergo oxidation or other post-translational modifications is analyzed, the glycans represent typical structures, with less than about 200 different glycan species present, and glycopeptide assignments are *all* within ±2.5 ppm error, the HR-MS data should suffice for providing a high proportion of correct assignments. If this mass error criterion cannot be met, or if the protein has a large number of glycosylation sites or other post-translational modifications, or if a restricted database of glycans cannot be used to characterize the glycosylation, confirmatory data, such as MS/MS analyses, are required.

Acknowledgement

The authors thank the NIH for funding, grant number RO1GM077226.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ijms.2008.12.001.

References

[1] S. Banerjee, P. Vishwanath, J. Cui, D.J. Kelleher, R. Gilmore, P.W. Robbins, J. Samuelson, Proc. Natl. Acad. Sci. U.S.A. 104 (28) (2007) 11676.
[2] C. Liu, J.A. Dias, Biochem. Biophys. 329 (1996) 127.
[3] L.A. Bishop, T.V. Nguyen, P.R. Schofield, Endocrinology 136 (1995) 2635.
[4] M.J. Koury, Trends Biotechnol. 21 (11) (2003) 462.
[5] D.J. Harvey, Expert Rev. Proteomics 2 (2005) 87.
[6] Y. Mechref, M.V. Novotny, Chem. Rev. 102 (2002) 321.
[7] J. Zaia, Mass Spectrom. Rev. 23 (2004) 161.
[8] D.S. Dalpathado, H. Desaire, Analyst 133 (2008) 731.
[9] M. Wuhrer, M.I. Catalina, A.M. Deelder, C.H. Hokke, J. Chromatogr. B 849 (2007) 115.
[10] J. Zhao, W. Qiu, D.M. Simeon, D.M. Lubman, J. Proteome Res. 6 (2007) 1126.
[11] E.P. Go, J. Irungu, Y. Zhang, D.S. Dalpathado, H.-X. Liao, L.L. Sutherland, S.M. Alam, B.F. Haynes, H. Desaire, J. Proteome Res. 7 (4) (2008) 1660.
[12] D.S. Dalpathado, J. Irungu, E.P. Go, K. Nortan, G.R. Bousfield, H. Desaire, Biochemistry 45 (2006) 8665.
[13] M.R. Flack, J. Froehlich, A.P. Bennet, J. Anasti, B.C. Nisula, J. Biol. Chem. 269 (1994) 14015.
[14] F.M. Valove, C. Finch, J.N. Anasti, J. Froehlich, M.R. Flack, Endocrinology 135 (1994) 2657.
[15] L.A. Bishop, D.M. Robertson, N. Cahir, P.R. Scholfield, Mol. Endocrinol. 8 (1994) 722.
[16] T. Saneyoshi, K. Min, X.J. Ma, Y. Nambo, T. Hiyama, S. Tanaka, K. Shiota, Biol. Reprod. 65 (2001) 1686.
[17] D. Goldberg, M. Bern, S. Parry, M. Sutton-Smith, M. Panico, H.R. Morris, A. Dell, J. Proteome Res. 6 (10) (2007) 3995.
[18] O. Krokhin, W. Ens, K.G. Standing, J. Wilkins, H. Perreault, Rapid Commun. Mass Spectrom. 18 (2004) 2020.
[19] T. Imre, G. Schlosser, G. Pocsfalvi, R. Siciliano, E. Molnar-Szollosi, T. Kremmer, A. Malorni, K. Vekey, J. Mass Spectrom. 40 (2005) 1472.
[20] J. Irungu, D.S. Dalpathado, E.P. Go, H. Jiang, H.V. Ha, G.R. Bousfield, H. Desaire, Anal. Chem. 78 (2006) 1181.
[21] Y. Zhang, E.P. Go, H. Desaire, Anal. Chem. 80 (9) (2008) 3144.
[22] H.X. Liao, L.L. Sutherland, S.M. Xia, M.E. Brock, R.M. Scearce, S. Vanleeuwen, S.M. Alam, M. McAdams, E.A. Weaver, Z. Camacho, B.J. Ma, Y. Li, J.M. Decker, G.J. Nabel, D.C. Montefiori, B.H. Hahn, B.T. Korber, F. Gao, B.F. Haynes, Virology 353 (2006) 268.
[23] E.P. Go, K.R. Rebecchi, D.S. Dalpathado, M.L. Bandu, Y. Zhang, H. Desaire, Anal. Chem. 79 (2007) 1708.
[24] J. Irungu, E.P. Go, Y. Zhang, D.S. Dalpathado, H.X. Liao, B.F. Haynes, H. Desaire, J. Am. Soc. Mass Spectrom 19 (2008) 1209.
[25] C.A. Cooper, E. Gasteiger, N.H. Packer, Proteomics 1 (2) (2001) 340.