A New Computer Program (GlycoX) To Determine Simultaneously the Glycosylation Sites and Oligosaccharide Heterogeneity of Glycoproteins

Hyun Joo An,¹ John S. Tillinghast,³ David L. Woodruff,§ David M. Rocke,³ and Carlito B. Lebrilla*², |¹

Department of Chemistry, Division of Biostatistics, Graduate School of Management, Biochemistry and Molecular Medicine, University of California, Davis, California 95616

Received June 16, 2006

A new computer program, GlycoX, was developed to aid in the determination of the glycosylation sites and oligosaccharide heterogeneity in glycoproteins. After digestion with the nonspecific protease, each glycan at a specific glycosylation site contains a small peptide tag that identifies the location of the glycan. GlycoX was developed in MATLAB requiring the entry of the exact masses of the glycopeptide and the glycan spectra in the form of a mass-intensity table and taking advantage of the accurate mass capability of the mass analyzer, in this case a Fourier transform ion cyclotron resonance (FT ICR) mass spectrometer. This program computes not only the glycosylation site but also the composition of the glycans at each site. Several glycoproteins were used to determine the efficacy of GlycoX. These glycoproteins range from the simple, with one site of glycosylation, to the more complex, with multiple (three) sites of glycosylation. The results obtained using the computer program were the same as those determined manually. Model glycoproteins yielded the correct results, and new glycoproteins with unknown glycosylation were examined with the site of glycosylation and the corresponding glycans determined. Furthermore, other functions in GlycoX, including an auto-isotope filter to identify monoisotopic peaks and an oligosaccharide calculator to obtain the oligosaccharide composition, are demonstrated.

Keywords: glycosylation site • oligosaccharide heterogeneity • software • oligosaccharide calculator • isotope filter • mass spectrometry

Introduction

Glycosylation is one of the most common forms of post-translational modification in eukaryotic proteins and is involved in many cell communication and signaling events.¹ Glycans play key roles in protein folding, cell–cell recognition, and the immune system.²,³ Defective or altered glycosylation can have profound biological implications and has been associated with a number of human diseases, for example, the congenital disorders of glycosylation⁴,⁵ and cancer.⁶

It has been estimated that at least 50% of human proteins are glycosylated.⁷ There are two major types, N- and O-glycosylation. The N-linked glycans are linked via an amide bond to an asparagine with a consensus sequence Asn-X-Ser (Thr), where X can be any amino acid except proline. N-linked oligosaccharides have a single core consisting of two N-acetylglucosamine (GlcNAc) and three mannoses (Man). The O-linked glycans are bound to either a serine or threonine with no single common core or consensus sequence at the attached peptide.

The determination of glycosylation sites and oligosaccharide heterogeneity is key toward understanding the specific biological roles of glycoproteins. Traditionally, site-specific glycosylation analysis has been extremely challenging due to the high complexity of glycoproteins. There have been several recent reports on glycosylation site analysis.⁸⁻¹⁶ Typical approaches to this task are based upon some combination of specific enzymatic proteolysis (usually with trypsin), fractionation of glycopeptides (most often by liquid chromatography or affinity chromatography), and glycopeptide analysis by mass spectrometry (MS).¹²⁻¹⁶ In some cases, deglycosylation of glycopeptides is concurrently performed with the information regarding the glycan discarded.¹²,¹⁶ Unfortunately, many glycoproteins are resistant to trypsin.¹⁷,¹⁸ Furthermore, the glycopeptides that are formed by specific enzymatic proteolysis are often too large for MS analysis because glycosylation may produce missed cleavages.¹⁴ For these reasons, the information regarding glycosylation is very often incomplete.

Recently, a new strategy for the rapid determination of N-glycosylation sites and site heterogeneity was introduced by
our laboratory.9 In this approach, a glycoprotein is subjected to nonspecific proteolysis with a highly active mixture of proteases known as Pronase. Nonglycosylated regions of the glycoprotein are digested to amino acids and dipeptides, while glycosylated regions are protected from protease activities by the oligosaccharides, which block the digestion of the associated peptide segment by steric interactions. The resulting mixture of glycopeptides is easily separated from the amino acids and salts using solid phase extraction (SPE) with porous graphitic carbon (PGC). In parallel, the glycoprotein is treated with PNGase F, an enzyme that specifically releases N-linked glycans from glycoproteins. However, this procedure is necessary only for large glycoproteins with several glycosylation sites. Smaller glycoproteins do not typically require the determination of the oligosaccharide constituents. In any case, the released glycans are also purified by SPE with PGC. The purified glyccopeptides and glycans are analyzed by matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry (MALDI–FTICR–MS), allowing the assignment of site-specific glycosylation and elucidation of glycan heterogeneity at individual sites.

To aid in the interpretation of this datum, we developed a unique and powerful software platform GlycoX that makes use of accurate masses belonging to the glycopeptides and glycans with the known protein sequence for assigning of N-linked and O-linked glycosylation with the additional determination of microheterogeneity. Although several tools such as GlycoMod,19 FindMod,20 and Glypeps21 are available for the determination of glycosylation sites from mass spectra, glycopeptides derived from specific proteases are required to predict the glycosylation sites. Recently, other groups have also utilized the approach using nonspecific or reduced-specificity proteolysis for glycoprotein analysis;22 however, there are no computer programs specifically developed to interpret automatic mass spectra obtained from nonspecific proteases such as Pronase E and protease K.

In this paper, we describe a new computer program (GlycoX) for determining site-specific glycosylation with microheterogeneity. The software described was developed in MATLAB and requires the entry of mass spectra as ASCII files. It takes advantage of the accurate mass capability of FTICR–MS. In addition, this program can be used to determine the glycosylation sites without the glycan information. GlycoX also includes an oligosaccharide calculator that accounts for multiple alkali addition to acidic oligosaccharides. To validate the program, several glycoproteins were characterized. These glycopolypeptides range from the simple, with one site of glycosylation, to the more complex, with at least three sites of glycosylation. It is also shown that GlycoX can be used to determine oligosaccharide composition of released glycans. An isotope filter is included to determine the monoisotopic quasimolecular ions in mixtures of glycopeptides or glycans readily.

**Experimental Section**

**General Experimental. 1. Pronase Digestion.** Pronase digestion has been described in great detail in an earlier publication.9 Briefly, Pronase E (~10 U) was added to the glycoprotein (~10 nmol) and incubated at 37 °C for 36–48 h. The reaction mixture was boiled to deactivate the enzyme. The digested glycoprotein was desalted and concentrated on a porous graphitic carbon (PGC) cartridge.

**2. Release of N-Linked Glycans by PNGase F.** The glycoprotein (~10 nmol) was dissolved in ammonium bicarbonate buffer solution (100 mM, pH 7.6). PNGase F (1 µL, ~10 U) was added, and the solution was incubated at 37 °C in a water bath for 17 h. The digestion mixture was boiled for 3 min to deactivate the enzyme. The digestion mixture was applied to PGC–SPE to purify the released glycans.

**3. Release of O-Linked Glycans by Reductive β-Elimination.** The glycoprotein was subjected to the reductive β-elimination. Alkaline borohydride solution (200 µL, mixture of 1.0 M sodium borohydride and 0.1 M sodium hydroxide) was added to 200 µg of glycoprotein. The mixture was incubated at 42 °C for 24 h. After the reaction period, a 1.0 M hydrochloric acid solution was slowly added with the reaction mixture in an ice bath to destroy excess sodium borohydride. The released glycans were purified by SPE employing PGC.

**4. Mass Spectrometry.** Mass spectra were recorded on an external source HiResMALDI (IonSpec corporation, Irvine, CA) equipped with a 7.0 T magnet and a pulsed YAG laser (355 nm) for ionization. 2,5-Dihydroxy benzoic acid (DHB) was used as a matrix (5 mg/100 µL in ethanol). A saturated solution of NaCl (or in some case KCl) in methanol was used as a dopant for the identification of the quasimolecular ion for either glycopeptides or oligosaccharides. The glyc peptide/oligosaccharide solution (1 µL) was applied to the MALDI probe followed by matrix solution (1 µL). The sample was dried under a stream of air prior to mass spectrometric analysis.

**5. Operating System.** The GlycoX program was written in MATLAB. We have used MATLAB 6 on several Windows platforms including 98, 2000, and XP.

**Subroutines in GlycoX.** The *ReadSpectrum* (*specfile*) subroutine reads spectral data from the file *specfile*. The user is asked to put quotes around the name of the file. The format of the specfile is ASCII, and the file contains two columns with masses and the intensities, for example,

<table>
<thead>
<tr>
<th>Mass (Da)</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500.010</td>
<td>72.1513</td>
</tr>
<tr>
<td>1500.0084</td>
<td>66.4353</td>
</tr>
<tr>
<td>1500.0158</td>
<td>60.8504</td>
</tr>
<tr>
<td>1500.0233</td>
<td>55.7041</td>
</tr>
<tr>
<td>1500.0307</td>
<td>51.2539</td>
</tr>
</tbody>
</table>

The output is a spectrum in GlycoX format—a matrix with two columns and many rows. The first column contains masses that are evenly spaced with separation delta (default value 0.01 Da). The second column has the intensities for the evenly spaced masses based on spline interpolation from the input data.

The *MonoGly*(spectrum) and *MonoPep*(spectrum) subroutines perform the isotope filter by taking the raw spectra and returning the estimate monoisotopic spectra in GlycoX format. The isotopic abundances of glycans and glycopeptides are slightly different and are accounted for in the respective subroutines (see next section).

The *SolveGlySpec*(spectrum, *modename*, *PercentMax*, *errtol*) performs the preprocessing by converting the raw spectrum and produces the resulting spectrum in the GlycoX format. *SolveGlySpec* runs *MonoGly* internally and should not be used on a monoisotopic spectrum. The input *modename* = ‘n+’, ‘o+’, ‘n-’, or ‘o-’ specifies the types of oligosaccharides with ‘n’ and ‘o’ representing N-linked and O-linked oligosaccharides, respectively, and ‘+’ and ‘-’ representing the positive and negative modes in MS. *PercentMax* is a threshold for identifying peaks. No peaks will be identified with height less than this percentage of the highest peak. *errtol* is the error allowed in parts per million of the mass accuracy (ppm). This routine also
runs GetPeakInfo and SolveGlyPeaks (see below) and displays predicted glycans (oligosaccharide composition).

The GetPeakInfo(xy, PercentMax) subroutine shows the peak information for any of the GlycoX format spectra. The xy input is a spectrum in GlycoX format (isotope-filtered or unfiltered). The input PercentMax sets the peak threshold. The subroutine returns PeakInfo, a matrix with two columns for each mass peak with mass and corresponding intensity.

The subroutine SolveGlyPeaks (PeakInfo, modename, errtol) determines the glycan composition based on the mass from exact masses of monosaccharide residues. PeakInfo is the file returned by GetPeakInfo. The files modename and errtol are the same as those used in SolveGlySpec. The subroutine prints an output table of possible solutions for the relevant peaks.

The subroutine SolveGlyFile (specfile, modename, PercentMax, errtol) solves the glycan composition as SolveGlyPeaks; however, the input can be the raw ASCII data (unfiltered). The output is the same as SolveGlyPeaks.

The subroutine SolveGP (GPInfo, GPModeName, GlyInfo, GlyModeName, SeqFile, errtol, MaxPepLen, glylist, GlyMat) determines the glycopeptides and the oligosaccharide heterogeneity. The subroutine prints output for masses and intensities given by GPInfo, a matrix just like PeakInfo in GetPeakInfo or SolveGlyPeaks, but for the glycopeptide mass values and intensities. GPModeName is ‘n+’, ‘o+’, ‘n-’, or ‘o-’, depending on which mode was used for the glycopeptide measurements. GlyInfo is the matrix of glycan mass values and intensities. GlyModeName is also ‘n+’, ‘o+’, ‘n-’, or ‘o-’ for the glycan measurements. SeqFile is a protein sequence file in FASTA format. errtol is again the error allowed, in parts per million. MaxPepLen sets the longest peptide length that GlycoX will consider. glylist is a list of the names of the glycans. The default is “glycans” from the routine StartGlycoX. GlyMat is a matrix with three rows: glycan masses on top, minimum counts below, and maximum counts below that. In StartGlycoX, we define NglyMat (N-linked glycan) and OGlyMat (O-linked glycan), but others can be created by the user.

Results and Discussion

The GlycoX program has three main functions, a ‘glycosylation site search’ for the determination of glycosylation sites with oligosaccharide heterogeneity in glycoproteins; an ‘oligosaccharide calculator’ for assigning oligosaccharide compositions for N-linked, O-linked, and chemically modified oligosaccharides; and an ‘auto-isotope filter’ for selecting the monoisotopic peaks from the mass spectra of glycopeptides, peptides, and glycans.  

1. Isotope Filter. An isotope filter was developed to simplify the mass spectra for analysis. Simplification is needed because isotopes will produce several peaks in the mass spectrum at masses m, m + 1, m + 2, and so forth. If all peaks are used, there can be interference and confusion between a higher isotopic peak of one molecule and the lower peaks of another. To have an unambiguous peak with each compound, we would prefer to work with a monoisotopic spectrum.  

For a given compound, the monoisotopic mass is the mass of the isotopic peak whose elemental composition is composed of the most abundant isotopes of those elements. Monoisotopic masses are used more often than average masses in mass spectrometry to relate peaks with their compositions. The GlycoX method uses the monoisotopic masses only and would give incorrect results if applied to the larger isotopic peaks. There is a need for a robust software to simplify the MS data because complex isotopic patterns make peak interpretation difficult.

Recently, a new algorithm, Iosconv, for deisotoping electrospray and MALDI mass spectra was developed by Hoffmann and co-workers. This algorithm was successfully used for peptides. However, there is currently no readily accessible software to filter the monoisotopic masses of other biomolecules such as glycans and glycopeptides. Different molecule classes need slightly different parameters for more efficient deisotoping algorithms (see below). Thus, isotope filter is a unique and efficient software for simplifying the MALDI–MS spectra of oligosaccharides, glycopeptides, and peptides.

An isotope filter can be used to deconvolute two overlapping quasimolecular ion signals. For example, shown in Figure 1a are two ionic species that differ by two mass units. The isotopic peaks of the lower mass interfere with the less abundant monoisotopic peaks of the larger mass. Performing the filter eliminates the contribution of the lower mass to the higher, although with some distortion, showing two distinct peaks in the region.

An algorithm that we call the isotope filter estimates the monoisotopic spectrum that would correspond to the measured spectrum with its isotopic peaks. For a given molecule,
there are rules for the relative abundances of isotopomer,

\[
y_{m+1} = ay_m
\]

\[
y_{m+2} = \left(\frac{a^2}{2} + b\right)y_m
\]

\[
y_{m+3} = \left(\frac{a^3}{6} + ab\right)y_m
\]

\[
y_{m+4} = \left(\frac{a^4}{24} + \frac{a^2b}{2} + \frac{b^2}{2}\right)y_m
\]

where,

\[
a = 0.00011n_H + 0.011n_C + 0.0036n_N
\]

is the expected number of single-isotope atoms per molecule \(^{12}\text{C}, \, ^2\text{H}, \, \text{and } ^{15}\text{N}\), and

\[
b = 0.0022n_O + 0.045n_S
\]

is the expected number of double-isotope atoms per molecule \(^{18}\text{O} \, \text{and } ^{34}\text{S}\). Figure 1b shows the spectrum for a glycan with elemental composition \(\text{C}_{112}\text{H}_{184}\text{O}_{80}\text{N}_8\). Fitted to the spectrum are predictions based on the rules for isotope ratios. The blue trace is the experimental plot. The green trace is based on the average composition of glycans of a mass 2944. The values for \(a\) and \(b\) are determined from an “average glycan” residue (“averagose”) analogous to “averagine” for peptides. The red trace is the synthesized spectrum with \(a\) and \(b\) calculated from the known glycan composition.

This can be written as a matrix relationship,

\[
y = \exp(aS + bS^2)y^M
\]

for

\[
y = \begin{bmatrix} y_m \\ y_{m+1} \\ y_{m+2} \\ \vdots \end{bmatrix}, \quad y^M = \begin{bmatrix} y_m \\ 0 \\ 0 \\ \vdots \end{bmatrix}
\]
and \( S \) is the shift matrix

\[
S = \begin{bmatrix}
0 & 0 & 0 & \cdots & 1 \\
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{bmatrix}
\]

that is, multiplying by \( S \) just shifts the whole spectrum up by one mass unit. The example shown here is for the simplest case: a spectrum of integer masses. The matrix \( S \) must always shift the spectrum by one whole mass unit; therefore in a case with precision for 100 data points per amu, the matrix would have ones along the 100th subdiagonal instead of the first.

For large enough masses, we can use an approximation

\[
a \approx \alpha m, \quad \text{and} \quad b \approx \beta m
\]

where \( \alpha \) and \( \beta \) are the average numbers of single and double isotopes per amu. Combining the relations for different molecules leads to

\[
y \approx \exp(\alpha m S + \beta m S^2) y^M
\]

where \( m \) is the diagonal matrix that multiplies each intensity by the mass, and \( S \) is the "shift" matrix that slides the spectrum down by one mass unit.

This means that

\[
y^M \approx \exp(-\alpha m S - \beta m S^2) y
\]

The GlycoX package has a very efficient way to compute this expression given \( y \).

For the molecule classes we have been working with, different parameters for the approximation of the peptides, glycans, and glycopeptides are used to give a more accurate output. We have found that the following coefficients provide a good approximation:

<table>
<thead>
<tr>
<th>molecule type</th>
<th>( A )</th>
<th>( B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>peptides</td>
<td>( 5.2 \times 10^{-4} )</td>
<td>( 3.0 \times 10^{-5} )</td>
</tr>
<tr>
<td>oligosaccharides</td>
<td>( 4.3 \times 10^{-4} )</td>
<td>( 6.1 \times 10^{-5} )</td>
</tr>
<tr>
<td>glycopeptides</td>
<td>( 4.6 \times 10^{-4} )</td>
<td>( 5.0 \times 10^{-5} )</td>
</tr>
</tbody>
</table>

Before-and-after pictures show the effect of using the filter. To illustrate the utility of the isotope filter, a complicated mixture of oligosaccharides from human tear was analyzed. Figure 2a shows the MALDI–MS spectrum before the "isotope filter" with several overlapping peaks. After the isotope filter (Figure 2b), the spectrum still contains many peaks, but is considerably reduced in complexity.

2. Oligosaccharide Calculator. GlycoX can be used to calculate the possible glycan compositions from experimentally determined masses. This is a stand-alone feature that can be used on the mass spectra of oligosaccharide mixtures. The user has the choice to define oligosaccharide type (N-linked or O-linked glycan), MS detection mode (positive or negative mode), and the minimum and maximum number of monosaccharides (Figure 3). Furthermore, all glycans types, that is, reduced glycans (alditol), unreduced glycans (aldehyde), and chemically derivatized glycans, can be calculated in GlycoX. A particular advantage of this program is the ability to easily input the sugar’s modification and the corresponding mass. To get biologically relevant oligosaccharide compositions, several restrictions were placed on the oligosaccharide calculator. First, the number of fucose residues should be less than or equal to the sum of the number of Hex plus HexNAc residues. Second, the N-linked glycans should have 2HexNAc and 3 Man corresponding to the core structure.

Oligosaccharides readily coordinate with Na\(^+\). Acidic oligosaccharides exchange acidic protons with Na\(^+\) to yield...
quasimolecular ions with compositions \([M - nH + (n+1)Na]^+\), where \(n\) = no. of acidic residues. We often observe multisodium adduct peaks in the mass spectra of acidic oligosaccharides. The Na\(^+\)-exchanged peaks yield multiple quasimolecular ions that often complicate MS spectrum. A few oligosaccharide calculator programs have recently been described, but there are no programs that account for multisodium-adducted peaks. GlycoX is unique because it shows all common possible sodium combinations with the oligosaccharides in order to calculate the acidic glycan composition.

N-linked glycans of bovine fetuin are highly glycosylated. To test the program, the oligosaccharides were released by PNGase F and analyzed by MALDI-MS in the positive ion mode. The resulting mass spectrum was entered into GlycoX to calculate the glycan compositions. Figure 4a shows the output page for the N-linked glycans of bovine fetuin. The results correctly determine the bi- and triantennary glycan compositions. All observed quasimolecular ions have \([M - nH + (n+1)Na]^+\), where \(n\) represents the number of sialic acid residues. Peaks corresponding to \([M - H + 2Na]^+\) of monosialylated biantenn-
nary and triantennary were observed in the mass spectrum at m/z 1976.69 and 2341.84, respectively. A disialylated triantennary composition were also found at m/z 2654.92 ([M - 2H + 3Na]+). A trisialylated triantennary composition was observed at m/z 2967.99 ([M - 3H + 4Na]+).

The composition of O-linked oligosaccharides can also be determined. In addition, the negative mode can also be analyzed as readily as the positive mode. The MS spectrum of an O-linked oligosaccharide was obtained in the negative ion mode and examined by GlycoX. The oligosaccharides were released from the egg jelly coats of *Xenopus tropicalis* by β-elimination. The output is shown in Figure 4b with all the oligosaccharides correctly determined. All of the peaks corresponded to the deprotonated species ([M - H]−).

3. Determination of Glycosylation Sites. The glycoproteins were digested with a nonspecific protease and purified as described previously.9 A portion of the glycoprotein sample was treated with PNGase F to release N-linked oligosaccharides. The digested glycopeptides and glycans were purified by PGC-SPE. The PGC fractions of the glycopeptides and the N-linked glycans were analyzed by MALDI-FTMS in both the positive and negative mode. The mass intensity (M/I) table, saved as

![Figure 5. An output page of N-linked glycosylation sites of XL CGL1 using (a) glycan and glycopeptide masses together and (b) only glycopeptide masses in GlycoX.](image-url)
Several glycoproteins were used to determine the efficacy of GlycoX. These glycoproteins range from the simple, with one site of glycosylation, to the more complex, with multiple sites of glycosylation. For comparison, the glycans and glycosylation sites were also determined manually using the same data. The results obtained using the GlycoX were identical to those determined manually. In addition, GlycoX found several more glycopeptides that were missed by the manual treatment. Model glycoproteins with known glycosylation sites such as ribonuclease B (with one glycosylation site) and chicken ovalbumin (with two sites and one occupied) were tested and ribonuclease B (with one glycosylation site) and chicken ovalbumin (with two sites and one occupied) were tested and the correct results (data not shown). The results obtained using the GlycoX were identical to those determined manually. In addition, GlycoX found several more glycopeptides that were missed by the manual treatment. Model glycoproteins with known glycosylation sites such as ribonuclease B (with one glycosylation site) and chicken ovalbumin (with two sites and one occupied) were tested and yielded the correct results (data not shown).

ASCII files, and the corresponding glycoprotein sequences from the Swiss-Prot/TrEMBL database saved in FASTA format (text file) were entered into GlycoX.

To determine the glycosylation sites and the accompanying glycans, the isotope filtered masses were then enumerated for combinations of glycan and peptides. Peptides of variable lengths, up to the user-specified lengths, were examined. The peptide composition was determined from the calculated mass by comparing all possible sequences from the monopeptide to the user-specified maximum peptide sequence. For glycan masses, the program generates possible combinations that can fit in the given mass. Alternatively, masses from the glycan profile spectra can be used. For complicated glycosylation, the measured glycan profile should be used.

Several glycoproteins were used to determine the efficacy of GlycoX. These glycoproteins range from the simple, with one site of glycosylation, to the more complex, with multiple sites of glycosylation. For comparison, the glycans and glycosylation sites were also determined manually using the same data. The results obtained using the GlycoX were identical to those determined manually. In addition, GlycoX found several more glycopeptides that were missed by the manual treatment. Model glycoproteins with known glycosylation sites such as ribonuclease B (with one glycosylation site) and chicken ovalbumin (with two sites and one occupied) were tested and yielded the correct results (data not shown).  

For examples of more complicated glycoproteins (with unknown glycosylation sites), cortical granule lectins (CGL1 and CGL2) from *Xenopus laevis* (XL) eggs and CGL from *X. tropicalis* (XT) eggs were examined. An output page of N-linked glycosylation of CGL2 of *X. laevis* using only glycopeptides masses were examined by GlycoX. As expected, one obtains more glycopeptide hits with the lower tolerance. For this example, 1836 hits were obtained with a mass tolerance of 1.0 Da, 502 with 0.1 Da, and 181 with 0.05 Da. The number of hits decreases further to 42 with a mass tolerance of 0.02 Da. Upon inspection, we found that all 42 were correct and all other hits above this number were false positive. However, when glycan information is provided, then the mass accuracy can be relaxed to 0.1 Da. In this case, 37 hits were obtained corresponding to "true" glycopeptides. These results confirm that high mass accuracy (<10 ppm) is needed for the determination of glycosylation sites when there is no glycan information. However, if some glycan information is known, the false-positive hits can be decreased even with less mass accuracy. These experiments can therefore be performed on high mass accuracy instruments such as orbitraps and time-of-flight, and maybe even ion traps or quadrupoles, when glycan information is available.

The CGL1 of *X. laevis* has two potential glycosylation sites. On the basis of this analysis, we found that both sites are occupied (154Asn and 163Asn). The oligosaccharides associated...
with each sites are shown in cartoon form in Figure 6. The putative glycan structures in Figure 6 were based solely on the glycan masses and were not verified with additional analysis. However, these assignments are consistent with several structures as determined independently by NMR. The putative glycan structures in Figure 6 were based solely on the glycan masses and were not verified with additional analysis. However, these assignments are consistent with several structures as determined independently by NMR. 34 CGL2 of X. laevis, with three potential sites, was also examined and was found to contain only two occupied sites (154Asn and 217Asn). The glycosylation sites of this protein have been determined manually in the previous publication. 9 The GlycoX analysis yields the same exact results.

Conclusions

The program GlycoX aids in the interpretation of MS data from a nonspecific protease treatment of a glycoprotein. The determination includes both the occupation of the site and site heterogeneity. We employed primarily FTICR–MS for the analysis; however, similar analyses can be performed on other high-mass accuracy instruments that are now becoming widely available. Additionally, both ionization methods, MALDI and electrospray ionization (ESI), can be used for this analysis. The method of nonspecific protease digestion with the automated data analysis could find wide applications in the determination of glycosylation sites on glycoproteins. The current method of tryptic digestion has specific limitations, but it provides well-defined peptide chains. For this reason, the two methods may be complementary.

The software will be made available upon requests. To fully automate the analysis, it would be desirable to have a graphical interface that draws the glycans and the associated sites automatically.

Acknowledgment. We gratefully acknowledge the financial support provided to C.B.L. by the National Institutes of Healthy (ROI GM049077) and to D.M.R. by the National Cancer Institute (P30 CA093373-04) and the National Institute of Environmental Health Sciences (P42-ES0469). We also thank Drs. Jerry Hedrick and Thomas Peavy for providing the sample of Xenopus cortical granule lectin.

References

(1) Varaki, A. Glycobiology 1993, 3, 97–130.
(34) Song, Y. W., N. J.; Shimoda, Y.; Hedrick, J. L. unpublished results.

PR0602949