Affinity Enrichment and Characterization of Mucin Core-1 Type Glycopeptides from Bovine Serum

Zsuzsanna Darula‡ and Katalin F. Medzihradszky‡§¶

The lack of consensus sequence, common core structure, and universal endoglycosidase for the release of O-linked oligosaccharides makes O-glycosylation more difficult to tackle than N-glycosylation. Structural elucidation by mass spectrometry is usually inconclusive as the CID spectra of most glycopeptides are dominated by carbohydrate-related fragments, preventing peptide identification. In addition, O-linked structures also undergo a gas-phase rearrangement reaction, which eliminates the sugar without leaving a telltale sign at its former attachment site. In the present study we report the enrichment and mass spectrometric analysis of proteins from bovine serum bearing Gal[1–3GalNAcO (mucin core-1 type) structures and the analysis of O-linked glycopeptides utilizing electron transfer dissociation and high resolution, high mass accuracy precursor ion measurements. Electron transfer dissociation (ETD) analysis of intact glycopeptides provided sufficient information for the identification of several glycosylation sites. However, glycopeptides frequently feature precursor ions of low charge density (m/z > ~850) that will not undergo efficient ETD fragmentation. Exoglycosidase digestion was utilized to reduce the mass of the molecules while retaining their charge. ETD analysis of species modified by a single GalNAc at each site was significantly more successful in the characterization of multiply modified molecules. We report the unambiguous identification of 21 novel glycosylation sites. We also detail the limitations of the enrichment method as well as the ETD analysis. Molecular & Cellular Proteomics 8:2515–2526, 2009.

Glycosylation is among the most prevalent post-translational modifications of proteins; it is estimated that over half of all proteins undergo glycosylation during their lifespan (1). Glycosylation of secreted proteins and the extracellular part of membrane proteins occurs in the endoplasmic reticulum and the contiguous Golgi complex. The side chains of Trp, Asn, and Thr/Ser residues can be modified, termed as C-, N-, and O-glycosylation, respectively (2, 3). In addition, O-glycosylation also occurs within the nucleus and the cytosol: a single GlcNAc residue modifies Ser and Thr residues. O-GlcNAc glycosylation fulfills a regulatory/signaling function similar to phosphorylation (4).

From an analytical point of view, C-glycosylation is the simplest. A consensus sequence has been defined: WXXW where the first Trp is modified, and the modification, a Man moiety, readily survives sample preparation and mass spectrometric analysis, including collisional activation (5). Investigation of N-glycosylation is also facilitated by several factors. First, N-glycosylation again has a well defined consensus sequence: NX(S/T/C) where the middle amino acid cannot be Pro (6). Second, there is a universal core glycan structure: GlcNAc₂Man₃; and this core is conserved across species. Third, a specific endoglycosidase, peptide N-glycosidase F, has been identified. This enzyme cleaves the carbohydrate structure from the peptide, leaving behind a diagnostic sign: the Asn residue is hydrolyzed to Asp, inducing a mass shift of +1 Da. By contrast, analysis of O-glycosylation is hampered by a lack of (i) a consensus sequence, (ii) a universal core structure, and (iii) a universal endoglycosidase or gentle chemical hydrolysis method to facilitate analysis.

Glycosylation shows a high degree of species and tissue specificity; the same site may be modified by a wide variety of different glycan structures, and unmodified variants of the protein may occur simultaneously (7–9). Disease and physiological changes also may alter the glycosylation pattern (10–12). The biological role(s) of glycosylation has been studied extensively (13–15), although such studies are seriously hampered by the difficulties of glycosylation analysis.

Most secreted proteins are glycosylated; and thus, mammalian serum is rich in glycoproteins. On the other hand, O-linked glycoproteins represent a small percentage of the serum protein content. Glycoproteins may display a befuddling heterogeneity both in site specificity and site occupancy. Thus, the enrichment of modified proteins or peptides is necessary for their characterization, and different techniques have been tested for this purpose. Lectin affinity chromatography is a popular method for selective isolation of glycoproteins and glycopeptides. Concanavalin A can be
O-Linked Glycopeptides

used to isolate oligomannose type glycopeptides (16), wheat germ agglutinin is applied for GlcNAc-containing compounds (16, 17), and jacalin is selective for core-1 type O-glycopeptides (18, 19). Lectins with preferential affinity for fucosylated and sialylated structures can also be utilized (12). Non-selective capture of glycopeptides can be performed using hydrophilic interaction chromatography (20, 21) or size exclusion chromatography (22). A recent approach applies porous graphite columns for semiselective enrichment (23), whereas the acidic character of sialylated glycopeptides has also been exploited via titanium dioxide-mediated enrichment (24). Finally vicinal cis-diols can be selectively captured using boronic acid derivatives (25–27). All methods described here provide some glycopeptide enrichment from non-glycosylated peptide background, but all also suffer from significant non-selective binding. N-Linked glycoproteins may also be selectively captured on hydrazide resin following periodate oxidation (28). This approach requires enzymatic deglycosylation to release the captured peptides for analysis, therefore excluding the determination of the carbohydrate structure.

Intact glycopeptide characterization still represents a significant challenge. Edman degradation, either alone or in combination with mass spectrometry, has been utilized for such tasks (29, 30). CID analysis of O-linked glycopeptides has limited utility. (i) MS/MS analysis cannot differentiate between the isomeric carbohydrate units and usually does not reveal the linkage positions and the configuration of the glycosidic bonds. (ii) Such spectra are typically dominated by abundant product ions associated with carbohydrate fragmentation, namely non-reducing end oxonium ions and product ions formed via sequential neutral losses of sugar residues from the precursor ions. (iii) The glycan is cleaved from the peptide via a gas-phase rearrangement reaction, and as a result the peptide itself and most peptide fragments (if any) are detected partially or completely deglycosylated (31–33). Recently a different approach, the combination of positive and negative ion mode infrared multiphoton dissociation, was found to provide conclusive structural assignment for some O-linked glycopeptides (34).

However, two novel MS/MS techniques, electron capture dissociation (ECD),1 which is performed in FT-ICR mass spectrometers (35), and electron transfer dissociation (ETD), which is performed in various ion trapping devices (36), may represent the real breakthrough. In both cases an electron is transferred to multiply protonated peptide cations, triggering peptide fragmentation at the covalent bond between the amino group and the α-carbon, producing mostly c and radical z+ product ions while leaving the side chains intact.

ETD is typically more efficient than ECD and thus leads to more comprehensive fragmentation. In addition, ETD can be performed in ion traps and thus, at a higher sensitivity level, especially in a linear ion trap. Because it has been observed that there are instances when the electron transfer is efficient and still no significant fragmentation occurs, ETD is usually combined with supplementary (and gentle) CID activation (37). O-Glycosylation analysis using these new dissociative techniques has been investigated (38, 39). However, because of the complexity of extracellular O-glycosylation, analysis of complex mixtures is rarely attempted (18), and the above techniques are usually restricted to the analysis of purified proteins.

In this study we present the analysis of secreted O-linked glycopeptides. Lectin (jacalin) affinity chromatography was used to achieve some enrichment of core-1 O-GalNAcα type carbohydrate-carrying glycopeptides from bovine serum. The glycopeptide fractions were subjected to CID and ETD analysis. These experiments were performed on a linear ion trap-Orbitrap hybrid mass spectrometer (40). The Orbitrap delivered high resolution, high mass accuracy for the precursor ions, whereas the linear trap provided high sensitivity MS/MS analyses. Some fractions were also subjected to sequential exoglycosidase digestions, and glycopeptides retaining only the proximal GalNAc residues were analyzed. ProteinProspector v5.2.1, developed to accommodate ETD product ion spectra, aided data interpretation (41). We identified 26 glycosylation sites from bovine serum unambiguously; 21 of these sites have never been reported by any other study. No other single study to date has yielded so much information about O-linked glycosylation sites.

EXPERIMENTAL PROCEDURES

Jacalin Affinity Chromatography

Affinity columns were prepared by packing agarose-bound jacalin (Vector Laboratories, AL-1153; binding capacity, >4 mg of asialofetuin/ml of gel) into perfluoroalkylalkane tubing (Upchurch Scientific, 1507L) equipped with a 0.5-μm frit at its distal end. The pressure was maintained at ~4 megapascals during packing. Two affinity columns were prepared: column 1, 1 × 2000 mm (CV, 1.57 ml); and column 2, 1 × 200 mm (CV, 0.157 ml).

Affinity enrichment was performed on an Amersham Biosciences AKTA FPLC system (complete with P920 pump, M925 gradient mixer, UPC-900 UV conductivity monitor, and Frac 900 fraction collector) equipped with a 2-ml sample loop. After introducing the sample (flow rate, 50 μl/min), the column was washed with solvent A (175 mM Tris-HCl, pH 7.5; flow rate, 150 μl/min), and then the species bound were eluted with solvent B (0.8 M galactose, 175 mM Tris-HCl, pH 7.5; flow rate, 150 μl/min), collecting 2-min fractions.

Single Enrichment—Tryptic peptides from fetal bovine serum were fractionated on column 1 (washing with 8 CVs of solvent A and elution with 6 CVs of solvent B). Fractions of interest were acidified and desalted on C18 reversed phase (Varian Omnix A5703100 100-μl pipette tip) prior to LC/MS analysis.

Double Enrichment—Fetal bovine serum was directly injected to column 1 (washing with 8 CVs of solvent A and elution with 6 CVs of...
solvent B. Combined fractions of interest were digested with trypsin and subjected to a second enrichment on column 2 (washing with 20 CVs of solvent A and elution with 30 CVs of solvent B). Fractions of interest were acidified and desalted on C18 reversed phase (Varian Omix A57003100 100-μl pipette tip) prior to LC/MS analysis.

Tryptic Digestion

Samples were complemented with guanidine hydrochloride to give a final concentration of 6 M. Disulfide bridges were reduced using DTT (56 °C for 60 min), and the resultant free sulfhydryl groups were derivatized using iodoacetamide (1.1 x equivalent to DTT; 60 min in the dark at room temperature). Samples were then diluted 8-fold with 100 mM ammonium bicarbonate to reduce the guanidine hydrochloride concentration and incubated with porcine trypsin (Fluka, 95614; 1–2% (w/w) of the estimated protein content) at 37 °C for 6 h. Digestion was stopped by adding acetic acid (final pH 3). The resulting peptide mixtures were desalted on a C18 reversed phase column (Vydac 218TP52205), lyophilized, and stored at −20 °C prior to analysis.

Partial Deglycosylation of O-Glycopeptides

Sialic acid and β-galactose units of glycopeptides were removed by incubation with neuraminidase (New England Biolabs, P0720; 50 units in 100 mM sodium citrate, pH 6.0) for 1 h at 37 °C followed by overnight treatment with β-galactosidase (New England Biolabs, P0726; 10 units in 100 mM sodium citrate, pH 4.5) at 37 °C. Enzymatic deglycosylation was stopped by acidification to pH 3 with formic acid, and the resulting peptide mixtures were desalted on C18 reversed phase (Millipore, ZTC18S960; 10-μl C18 pipette tip).

Mass Spectrometry

Glycopeptide mixtures were separated on a nanoflow reversed phase HPLC system directing the eluent to nonospray sources of ESI-MS instruments operating in positive ion mode. The operating parameters for the different LC/MS systems were as follows. A quadrupole-orthogonal time-of-flight hybrid mass spectrometer (Q-TOF Premier) was on line-coupled with a nano-HPLC system (nanoACQUITY) (both Waters Micromass). For LC, in-line trapping onto a nanoACQUITY UPLC trapping column (Symmetry, C18 5 μm, 180 μm x 20 mm) (15 μl/min with 3% solvent B) followed by a linear gradient of solvent B (10–50% in 40 min; flow rate, 250 nl/min; nanoACQUITY UPLC BEH C18 column, 1.7 μm, 75 μm x 200 mm) was used. Solvent A was 0.1% formic acid in water, and solvent B was 0.1% formic acid in acetonitrile. For MS, CID spectra were acquired of the most abundant multiply charged ion from each MS survey with collision energy adjusted automatically according to the charge state and m/z value of the ion selected. Dynamic exclusion was also enabled; exclusion time was 60 s.

The linear ion trap-Orbitrap (LTQ-Orbitrap, Thermo Fisher Scientific) was on-line-coupled to a nanoACQUITY HPLC system. Reversed phase chromatography was performed using the same column and the same solvents as described above. Peptides were eluted by a gradient from 2 to 35% solvent B in 35 min followed by a short wash at 50% solvent B before returning to starting conditions. For MS, data acquisition was carried out in data-dependent fashion acquiring sequential CID and ETD spectra of the three most intense, multiply charged precursor ions identified from each MS survey scan. (MS spectra were acquired in the Orbitrap, and CID and ETD spectra were acquired in the linear ion trap.) Ion populations within the trapping instruments were controlled by integrated automatic gain control (AGC). For CID, the AGC target was set to 30,000 with dissociation at 35% of normalized collision energy; activation time was 30 ms. For ETD, the AGC target values were set to 30,000 and 200,000 for the isolated precursor cations and fluoranthene anions, respectively, allowing 100 ms of ion/ion reaction time. Supplemental activation for the ETD experiments was enabled. Dynamic exclusion was also enabled; exclusion time was 60 s.

Data Interpretation

Data from the Q-TOF experiments were manually evaluated and were mostly used to ascertain that the samples contained a significant number of glycopeptides. Peak lists from LTQ-Orbitrap raw data files were created by using the University of California San Francisco in-house peak picking program PAVA (42) as well as Bioworks 3.3.1 SP1. Both softwares generate separated CID and ETD peak lists. Database searching was performed by ProteinProspector v.5.2.1 against the Swiss-Prot database (April 24, 2008), which was supplemented with a random sequence for each entry, and the species was specified as Bos taurus (10170/725568 entries searched). Search parameters were as follows. Trypsin was selected as the enzyme, one missed cleavage was permitted, and nonspecific cleavage was also permitted at one of the peptide termini. The mass accuracy considered was 15 ppm for precursor ions and 0.6 Da for the fragment ions. The fixed modifications were carbamidomethylation of Cys residues. Variable modifications were the default modifications, i.e. the acetylation of protein N termini, Met oxidation, and the cyclization of N-terminal Gin residues, supplemented with a HexHexNac or SAHexHexNac carbohydrate modification on Thr and Ser residues. A maximum of two modifications per peptide was permitted. The same search parameters were used for the ETD data after the exoglycosidase digestion except Ser and Thr residues were considered modified by HexNac only, and three modifications per peptide were permitted. Search parameters for CID data acquired after the exoglycosidase treatment also included a modification of 203–203.1 Da on Ser and Thr residues leading to neutral losses of the same mass value; i.e. fragments were assumed to be unmodified. For all reported results the acceptance criteria were as follows: minimum peptide score, 15; minimum protein score, 15. All glycopeptide identifications having a maximal expectation value of 0.5 were manually inspected. We repeated our searches considering two missed cleavages as well as non-tryptic cleavages at both peptide termini and included some glycopeptides from these additional results.

RESULTS AND DISCUSSION

Enrichment of O-glycosylated peptides from bovine serum was achieved by affinity chromatography utilizing jacalin-agarose. Jacalin, a plant lectin from Artocarpus integrifolia, binds GalNACα that is unsubstituted at C-6 (19, 43). We briefly validated the binding using a standard, fetuin trypptic digest, and then moved onto a more complex sample, the trypptic digest of bovine serum. It was found that O-linked glycopeptides indeed were enriched; however, a very high background due to nonspecific binding was observed. In the hope of eliminating the nonspecific background we also performed a two-step enrichment. First glycoproteins were enriched, and then the trypptic digest of the retained fractions was subjected to a second enrichment step. The result of the protein level enrichment is shown in Fig. 1. The retained fraction accounts for ~5% of the total UV absorbance. This relatively high amount can be accounted for by considering the high con-
O-Linked Glycopeptides

Fig. 1. UV absorbance chromatogram (280 nm) of fetal bovine serum-derived proteins subjected to jacalin affinity chromatography. The Gal-containing solution was added as indicated. Approximately 5% of total protein content was retained. Two-minute fractions were collected. Fractions 22–24 were combined for further analyses. mAU, milliabsorbance units.

centration of fetuin in fetal serum (44) but also reflects non-specific binding to the affinity column.

To confirm that glycopeptide enrichment indeed occurred, the resulting glycopeptide-containing fractions were subjected to a quick (1-h) data-dependent LC/MS/MS analysis on a Q-TOF instrument. Data were screened for characteristic glycopeptide fragments (31–33), for example for oxonium ions at m/z 204 (GalNAc) or m/z 292 (SA) as well as for sequential sugar losses from the precursor ion: −291 Da (SA), −162 Da (Gal), and −203 Da (GalNAc). Carbohydrate assignments were based on the lectin specificity, supported by exoglycosidase digestions (see later text). One of the double enrichment experiments was completely evaluated manually, and CID data indicated the presence of ~130 glycopeptides. Unfortunately these spectra hardly ever contain sufficient information for the identification of the glycopeptides.

To gain sufficient information to determine both the identity of the glycopeptides and their site(s) of modification, a technique that breaks the peptide backbone while leaving the modification(s) intact is required. Fortunately ETD offers precisely these features.

Thus, glycopeptide-containing fractions were subjected to LC/MS/MS experiments on a linear ion trap-Orbitrap hybrid instrument (LTQ-Orbitrap). Mass measurement was performed in the Orbitrap with high resolution and mass accuracy. Multiply charged precursor ions were subjected to both CID and ETD analysis in the linear ion trap. All spectra presented in this study were acquired in these analyses. CID spectra acquired in the linear ion trap were even less informative than those acquired in the Q-TOF instrument; they displayed practically exclusively carbohydrate fragments (oxonium ions as well as glycan-related neutral losses). If the charge state of the precursor ion was higher than 2, such fragment ions were frequently observed in multiple charge states (Fig. 2 and Scheme 1). In summary, low energy CID fragmentation of these glycopeptides provided sufficient information to confirm the presence of a glycopeptide, to determine the size and the number of the sugar units, and to ascertain the mass of the unmodified peptide.

Database searches with the ETD data were performed using ProteinProspector v5.2.1. Two different glycans were considered on Ser and Thr residues: the GalGalNAc disaccharide and its sialylated version. The fractions contained unmodified peptides from a wide variety of serum proteins, and a significant portion of these displayed non-tryptic termini due to proteolytic enzyme activity in the serum (data not shown) similar to glycopeptides identified from these experiments. Half of the glycopeptides listed in Tables I and II feature
non-tryptic termini. (For further examples see supplemental Tables 1 and 2.) Manual validation of the automated assignments led to the unambiguous identification of 17 glycosylation sites (listed with the best representative glycopeptide sequences in Table I; all automated assignments are presented in supplemental Table 1). Although ETD fragmentation is considered to leave the amino acid side chains intact, the glycopeptides investigated herein still exhibited a degree of carbohydrate fragmentation from the charge-reduced species (see Fig. 3, Scheme 2, and supplemental figures), which was unaccounted for within the parameters of the database search algorithm. Thus the score would be lowered by the presence of prominent, unassigned product ions. This fragmentation could result from the supplemental activation collisional warming, but radical site-induced glycosidic bond fragmentation has been reported previously in ECD experiments (45).

An additional problem is the reliability of site assignment(s). Database searching algorithms routinely assign a modification site to sequence data even where unambiguous assignment information is unavailable. This is a general computational problem. O-Linked glycopeptides frequently feature Pro residues (see below), which may form y type product ions following the supplemental activation step, but no ETD-induced fragmentation occurs N-terminal to this imino acid residue. O-Linked glycopeptides also frequently feature multiple adjacent hydroxy amino acids, and unless near complete sequence coverage is achieved, the modification site cannot be unambiguously identified. Plenty of examples are listed in the supplementary tables, and supplemental Figs. 13, 21, and 28 illustrate these problems.

Only a small portion of the total glycopeptide content of the sample was fully characterized from these data. Unfortunately ETD fragmentation of precursor ions at $m/z \sim 850$ was inefficient. In the majority of the product ion spectra generated for these species only the presence of the charge-reduced precursor ions indicated that electron transfer had indeed occurred. This adverse effect of low charge density on peptide fragmentation has been documented (37).

Although we cannot control the extent of ETD fragmentation of any given ion, one could try to increase the charge density to overcome poor fragmentation efficiency. This
O-Linked Glycopeptides

Glycosylation sites identified from the ETD data of intact glycopeptides

Glycosylation sites identified and the carbohydrate structure detected are printed in bold. Residues in parentheses represent the amino acid residues before and after the sequence stretch identified. They are included to demonstrate the extent of nonspecific cleavages.

<table>
<thead>
<tr>
<th>Residues before and after the sequence stretch identified. They are included to demonstrate the extent of nonspecific cleavages.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residues in parentheses represent the amino acid residues before and after the sequence stretch identified. They are included to demonstrate the extent of nonspecific cleavages.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Glycosylation sites identified from CID spectra after exoglycosidase treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycosylation sites identified and the carbohydrate structure detected are printed in bold. Residues in parentheses represent the amino acid residues before and after the sequence stretch identified. They are included to demonstrate the extent of nonspecific cleavages.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycosylation sites identified from the ETD data of intact glycopeptides</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Glycosylation sites identified from the ETD data of intact glycopeptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycosylation sites identified and the carbohydrate structure detected are printed in bold. Residues in parentheses represent the amino acid residues before and after the sequence stretch identified. They are included to demonstrate the extent of nonspecific cleavages.</td>
</tr>
</tbody>
</table>
transport (apolipoproteins and fetuins), enzymes or enzyme inhibitors (pantetheinase, inter-\(\alpha\)-trypsin inhibitor, and plasma serine protease inhibitor), cell signaling factors (insulin-like growth factor II), and proteins involved in blood coagulation (kininogen-1, fibrinopeptide, and fibronectin). None of these are low level proteins in serum. However, a rough estimation is that glycosylation sites identified in the present study belong to proteins that are 2–3 orders of magnitude less abundant than serum albumin in fetal serum.

Prediction tools especially tailored to mucin type O-linked glycosylation have been developed previously. For instance, NetOGlyc (51) considers primary sequence, surface accessibility, secondary structure information, and distance constraints for prediction of glycosylation. This software predicted 10 of the 26 sites identified unambiguously in the present study. The CKSAAP-OglySite software (52) predicted correctly 7 of the 26 identified glycosylation sites with higher false positive rates compared with NetOGlyc.

The first step in O-glycan synthesis is the transfer of a GalNAc unit from UDP-GalNAc to the protein. At least 21 enzymes, polypeptide N-acetylgalactosaminyltransferases are known to be able to perform this job (2). All these enzymes utilize the same donor substance but may display differential preferences for protein substrates. Thus, it is not surprising that no consensus sequence(s) for O-glycosylation has been determined. Nonetheless there have been attempts to identify sequence motifs for O-glycosylation (53). There are structural features that seem to be preferential for this modification. Most notably, modification sites are located in Ser/Thr-rich surroundings featuring Pro residues. There is speculation that the Pro residues help to expose the Ser/Thr residues for efficient glycosylation. We tested these predictions with the glycosylation sites determined in our study. To uncover any patterns in preferential O-glycosylation within our data set, a symmetrical window of 16 flanking amino acids surrounding the modification sites was interrogated. The frequency of occurrence of these amino acids within flanking sequences was compared with that of the precursor proteins. Increased observation of Pro residues was noted around modification sites with a proline over-representation factor of 2.7 (16.4% frequency around glycosylation sites compared with 6.2% in the full protein sequences). A small over-representation of the combined Ser and Thr content was also observed, but this increment was within one standard deviation of the average.

**Fig. 3.** ETD spectrum identifying a novel O-glycosylation site of bovine fetuin. The structure is HVGK(GalNAcGalSA)PIVGQPSIPGG-PVR. Neutral loss of sialic acid from the precursor ion was observed and is indicated as “-SA.”

**Scheme 2. Illustration of the fragmentation shown in Fig. 3.**
of our data. All these observations are in good agreement with previous results (53).

An interesting observation within our data set was that more than half of the glycosylation sites identified are located in the proximity of one of the termini of the protein. For example, the observed glycosylation sites in apolipoprotein C and pantetheinase precursor are both close to the C terminus. Although some of the identified sites seem to be somewhere in the middle of the unprocessed protein sequence, post-translational processing of signal and pre-sequences means that the location of the modification is in fact proximal to the terminus of the processed mature protein. For example, Thr99 and Thr106 of insulin-like growth factor II precursor (Swiss-Prot accession no. P07456) are located near the N terminus of the E-peptide. Similarly all the identified glycosylation sites in inter-H9251_H9251-trypsin inhibitor heavy chain H4 precursor (Q3T052) are in the middle of the translated sequence. However, the human homolog of this protein, Q14624, undergoes further processing to generate two separate inter-H9251-tryptsin inhibitor heavy chains, giving one 70-kDa and one 35-kDa protein. The bovine sequence shows close homology to its human counterpart (>70% identity); thus it is reasonable to assume that the bovine preprotein undergoes similar processing. Using the human sequence as a predictive template for the likely sites of cleavage in the mature bovine protein, the newly identified modification sites would lie near the N terminus of the shorter chain. Without a larger data set, we cannot yet make a comprehensive prediction of the likelihood of O-glycosylations to cluster adjacent to protein termini.

CONCLUSIONS: SHORTCOMINGS AND ADVANTAGES

Glycoproteins and glycopeptides have previously been enriched by lectin affinity chromatography. Jacalin has been used for larger scale O-glycopeptide enrichment in a prior study (18). However, in that study, the carbohydrates were removed prior to mass spectrometry analysis by oxidative elimination. Because no diagnostic signs remain at the previously modified residues after such deglycosylation, there is no direct evidence that the peptides identified had been glycosylated. In fact, based on our experience, the degree of nonspecific binding observed in jacalin enrichment is significant. Numerous unmodified peptides were identified in our study even after enrichment at both the protein and
peptide level. Moreover in a few cases the same sequences were detected both glycosylated and unmodified (fetuin (P12763) residues 334–348, 330–348, and 334–349). Thus, without direct detection of a carbohydrate-modified peptide, one cannot confirm that a glycopeptide was isolated by jacalin affinity chromatography.

Once the glycopeptides were isolated it became obvious that there is no simple paradigm for their full structural characterization. In most instances, CID analysis is only suitable for confirming the presence of glycopeptides. Although ion traps permit the automated setup of MS$^3$ experiments, one cannot suggest a universal linked scan approach for O-linked glycopeptide analysis because CID analysis of the most abundant MS/MS products will lead to further carbohydrate losses. ETD offers a more successful alternative. However, even with relatively small carbohydrate structures, the low charge density represents an obstacle for efficient ETD fragmentation. Limited fragmentation may prevent the site assignment or even identification of the glycopeptide. Nonspecific proteolytic cleavage due to proteolytic activity in serum and sugar fragmentation further confound data interpretation. There are some potential solutions, each of which is worthy of investigation. With shorter trypsinolysis time, i.e. with more missed cleavages, longer peptides may be created. The larger number of basic residues present in peptides thus generated will increase the charge density and thus improve the ETD fragmentation efficiency. There are two potential pitfalls for this approach. (i) Partial digestion may not improve the charge density, particularly in cases where glycosylation sites are clustered and the adjacent sequences are therefore modified. (ii) Most database searching algorithms do not consider fragment ions bearing charge states higher than 2$^+$/H$^+$; they will therefore not assign such ions, and thus the peptide matching will achieve a lower score and lower confidence. Using a hybrid LTQ-Orbitrap instrument, ETD fragments may also be measured at high resolution and subsequently processed to generate a deconvoluted peak list. With this approach, one has to consider that ion transmission of products from the ion trap to the Orbitrap currently leads to significantly decreased sensitivity. However, the newest LTQ-Orbitrap Velos may solve this problem (54).

The solution presented within this study, i.e. partial enzy-
matic removal of the oligosaccharide moieties, showed a strong improvement in the detection of multiply modified species. Even following partial deglycosylation, ETD fragmentation efficiency is still frequently insufficient to yield confident identification/site localization. A recent study has reported improved quality of ETD product ion spectra following proteolysis using an endoprotease cleaving selectively in front of Lys residues (55). This is an approach that will be further investigated for its general utility in glycopeptide analysis.

Where good quality mass spectrometric data were obtained, ProteinProspector v5.2.1 updated for ETD fragmentation, was used for database searching. We found that a 1.2 Da mass deviation had to be permitted for the product ions, to accommodate the alternate formation of \( z^{+1} \) and \( c^{-1} \) products (56). Consideration of the formation of these alternate fragments, a unique feature of this search engine, permitted lowering the mass tolerance to 0.6 Da. Introducing weighting factors for the different fragment ions also improved spectral assignments of ETD data (41). However, manual evaluation of the data was still essential. Development of a software package that could combine information obtained during the mass survey (permitting the identification of metal adducts), CID, and ETD experiments would be beneficial to this and other studies.

We present a practical and useful method for deciphering GalGalNAc core glycosylation from serum samples. More extensive fractionation of the samples and the application of different proteolytic enzymes will yield more, confidently assigned modification sites. A reliable catalog of such sites will undoubtedly lead to the development of more reliable modeling and prediction tools and eventually to better understanding of the biological function of the O-glycosylation of serum proteins.

<table>
<thead>
<tr>
<th>Identity</th>
<th>Name</th>
<th>Site&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Evidence&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>P01044</td>
<td>Kininogen-1 precursor</td>
<td>Thr&lt;sup&gt;136&lt;/sup&gt;</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ser&lt;sup&gt;581&lt;/sup&gt;</td>
<td>+</td>
</tr>
<tr>
<td>P02676</td>
<td>Fibrinogen ( \beta ) chain precursor (contains fibrinopeptide B)</td>
<td>Thr&lt;sup&gt;114&lt;/sup&gt;</td>
<td>+</td>
</tr>
<tr>
<td>P07456</td>
<td>Insulin-like growth factor II precursor</td>
<td>Ser&lt;sup&gt;95/97&lt;/sup&gt; or Thr&lt;sup&gt;99/99&lt;/sup&gt; (one of these 4 residues)</td>
<td>+</td>
</tr>
<tr>
<td>P07589</td>
<td>Fibronectin</td>
<td>Ser&lt;sup&gt;154&lt;/sup&gt;</td>
<td>+</td>
</tr>
<tr>
<td>P12763</td>
<td>( \alpha_2 )-HS-glycoprotein precursor</td>
<td>Thr&lt;sup&gt;163&lt;/sup&gt;</td>
<td>+ + +</td>
</tr>
<tr>
<td>P17690</td>
<td>( \beta_2 )-Glycoprotein 1 precursor</td>
<td>Thr&lt;sup&gt;33&lt;/sup&gt;</td>
<td>+</td>
</tr>
<tr>
<td>P19035</td>
<td>Apolipoprotein C-III precursor</td>
<td>Ser&lt;sup&gt;90&lt;/sup&gt; or Thr&lt;sup&gt;92&lt;/sup&gt;</td>
<td>+</td>
</tr>
<tr>
<td>Q0247</td>
<td>Apolipoprotein E precursor</td>
<td>Thr&lt;sup&gt;211&lt;/sup&gt;</td>
<td>+</td>
</tr>
<tr>
<td>Q0VCM5</td>
<td>Inter-( \alpha )-trypsin inhibitor heavy chain H1 precursor</td>
<td>Thr&lt;sup&gt;309&lt;/sup&gt;</td>
<td>+ + +</td>
</tr>
<tr>
<td>Q3T052</td>
<td>Inter-( \alpha )-trypsin inhibitor heavy chain H4 precursor</td>
<td>Thr&lt;sup&gt;705&lt;/sup&gt;</td>
<td>+ + +</td>
</tr>
<tr>
<td>Q58CQ9</td>
<td>Pantetheinase precursor</td>
<td>Ser&lt;sup&gt;10&lt;/sup&gt;</td>
<td>+ + +</td>
</tr>
<tr>
<td>Q58D62</td>
<td>Fetuin-B precursor</td>
<td>Ser&lt;sup&gt;296&lt;/sup&gt;</td>
<td>+ + +</td>
</tr>
</tbody>
</table>

<sup>a</sup> Normal text, sites previously known; italics, sites previously uncharacterized in bovine but reported in human; bold, new sites previously unreported.

<sup>b</sup> The modifying carbohydrate group is: ETD(3), GalNAcGalSA; ETD(2), GalNAcGal; ETD(1) and CID, GalNAc.

<sup>c</sup> See Ref. 46.

<sup>d</sup> Human protein, P01344, is listed as modified at Thr<sup>99</sup>, which may correspond to the same position in the bovine sequence (47).

<sup>e</sup> See Ref. 33.

<sup>f</sup> See Ref. 48.

<sup>g</sup> Human protein, P02656, has been reported modified at corresponding site Thr<sup>94</sup> (49).

<sup>h</sup> Human protein, P02649, is listed as modified at this conserved site (Thr<sup>212</sup>) in the Swiss-Prot database; no reference provided.

<sup>i</sup> Human protein, P19827, is listed as modified at Thr<sup>653</sup> (50), corresponding to bovine site reported.
REFERENCES


5. Hartmann, S., and Hofsteenge, J. (2000) Properdin, the positive regulator of complement, is highly C-mannosylated. J. Biol. Chem. 275, 28569–28574


O-Linked Glycopeptides


