ELSEVIER

Journal of Chromatography B, 752 (2001) 293–306

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Proteomics of glycoproteins based on affinity selection of glycopeptides from tryptic digests

Ming Geng, Xiang Zhang, Minou Bina, Fred Regnier*

Department of Chemistry, 3164A Brown Building, Purdue University, Lafayette, IN 47907-1393, USA

Abstract

Identification of glycoproteins in complex mixtures derived from either human blood serum or a cancer cell line was achieved in a process involving the steps of (1) reduction and alkylation, (2) proteolysis of all proteins in the mixture with trypsin, (3) affinity chromatographic selection of the glycopeptides with an immobilized lectin, (4) direct transfer of the glycopeptide fraction to a reversed-phase liquid chromatography (RPLC) column and further fractionation by gradient elution, (5) matrix-assisted laser desorption ionization mass spectrometry of individual fractions collected from the RPLC column, and (6) peptide identification based on a database search. The types of glycoproteins analyzed were; (1) *N*-type glycoproteins of known primary structure, (2) *N*-type glycoproteins of unknown structure, and (3) *O*-type glycoproteins glycosylated with a single *N*-acetylglucosamine. Identification of peptides from complex mixtures was greatly facilitated by either C-terminal sequencing with a carboxypeptidase mixture or by comparing chromatographic behavior and mass to standards, as in the case of a known protein. In addition, deglycosylation of peptides with N glycosidase F was necessary to identify *N*-type glycoproteins of unknown structure. The strength of this approach is that it is fast and targets specific molecular species or classes of glycoproteins for identification. The weakness is that it does not discriminate between glycoforms.

Keywords: Proteomics; Glycoproteins; Glycopeptides

1. Introduction

Proteomics focuses on the identification of large numbers of proteins from cellular extracts or biological fluids [1]. Samples normally contain thousands of protein species. The complexity of these biological materials and the difficulty in identifying proteins requires that separation steps be part of characterization. Separation by two-dimensional (2-D) gel electrophoresis is the most widely reported [2]. After electrophoresis, protein spots are excised from the gel and tryptic digested [3]. Proteolytic fragments are subsequently extracted from the gel and analyzed by mass spectrometry (MS) [4]. Protein identification is achieved by comparing the mass of tryptic peptides obtained from gel spots to masses of peptides derived from proteins in databases [5]. Each protein has a unique set of tryptic peptides that in composite provide a signature [6]. Peptides from other proteins may also be found when resolution in the 2-D gel is incomplete. Protein identification has also been achieved with single signature peptides [7]. Most proteins have multiple signature peptides.

Glycoproteins are problematic in the 2-D gel electrophoresis approach to proteomics [8]. Oligosaccharide heterogeneity often causes proteins of

*Corresponding author. Tel.: +1-765-4941-648; fax: +1-765-4940-359.

E-mail address: fregnier@purdue.edu (F. Regnier).

0378-4347/01/$ – see front matter
PII: S0378-4347(00)00550-8

identical primary structure to appear in multiple spots [9]. Glycoform resolution occurs because of differences in charge, often due to differences in sialic acid content. The presence of large numbers of isoforms complicates identification in several ways. One is that the concentration of protein in a spot is diminished when spread across many spots. A second is that the same protein will be identified many times. And finally, glycoforms contaminate adjacent spots on gels. Failure to resolve glycovariants is another problem. *O*-Glycosylation with *N*-acetylglucosamine or other neutral sugars at a serine or threonine residue is such a case. Glycoforms and the non-glycosylated parent can have the same charge [10]. Multiple site glycovariants of the same net charge would be another case. Methods that address these problems would be useful, particularly methods that examine changes in glycosylation patterns at particular sites.

Affinity chromatography affords another route to fractionation selectivity. Immobilized lectins have been used widely to select glycoproteins from biological extracts based on the types of glycosylation found in the protein [11]. Lectin columns have also been particularly useful in the selection and characterization of both tryptic glycopeptides and oligosaccharides derived from glycoproteins [12]. High mannose-type, complex-type, and hybrid-type sugar chains have been fractionated by using immobilized lectin columns in series, each with different binding affinity for the various glycotypes [13]. *Ricinus communis* lectin, *Galanthus nivalis* (snowdrop) lectin, wheat germ lectin, concanavalin A lectin, *Lens culinaris* (lentil) lectin, *Pisum sativum* (pea) lectin, *Vica fava* (fava) lectin, *Phaseolus vulgartis* erythroagglutinin, *Datuar stramonium* lectin, and *Tetracarpidium conophorum* (Nigerian walnut) lectin are some of the most widely used lectins. Among these, *Ricinus communis* lectin has the broadest selectivity [14]. Fractionation within a class can be accomplished by gradient elution of a specific column with increasing concentrations of a displacing agent [15]. Still higher resolution is achieved by fractionating affinity column fractions with reversed-phase liquid chromatography (RPLC), anion-exchange chromatography, or capillary electrophoresis [16]. Each of these separation methods probes a different property of the analyte. RPLC targets the hydrophobic character of a substance and is most useful with peptides. Anion-exchange chromatography under basic conditions probes both acidic character and the general structure of oligosaccharides. Capillary electrophoresis explores acidic character as well [17], but through the use of borate buffers it is possible to alter electrophoretic mobility by borate complex formation [18]. Borate ester formation is sensitive to the stereochemistry of vicinal diols in analytes. Most work on glycoconjugates has been directed at elucidating the structure and variability within the oligosaccharide portion of glycopeptides from a specific, purified protein [16]. Subsequent to selection and purification of glycopeptides containing the same peptide backbone, the oligosaccharide is cleaved from the glycoconjugate and further fractionated [16]. The peptide portion of the conjugate is discarded. Individual oligosaccharides obtained in this way have been sequenced chemically [19], by MS [20], or a combination of the two [21].

In contrast, it is the peptide portion of the conjugate that is of interest in proteomics, where it is used in the identification its parent. The 2-D gel electrophoresis approach to proteomics has been described above. Another strategy is to tryptic digest all the proteins in a mixture and after affinity selection of one, or a few peptides from each protein they are further fractionated and identified by MS [22]. Selection is generally directed at low abundance amino acids (i.e., histidine or cysteine) or some type of post-translational modification. Selection of glycopeptides from tryptic digests with lectins falls into this strategy [23]. The attractive feature of using glycopeptides to identify parent proteins is that the peptide portion of the molecule must have the mandatory glycosylation sequence [24]. Knowledge of mandatory sequence elements and amino acid composition greatly facilitate database searches. When the oligosaccharide structure is known, its mass can be subtracted from the mass of the observed peptide before searching databases. Peptides glycosylated with *N*-acetylglucosamine are an example [25]. The problem is when the mass of the oligosaccharide is unknown. The simple process of subtracting the mass of the oligosaccharide portion of the glycoconjugate from the total mass of the peptide cannot be used. It will be shown below that this problem can be addressed by removal of the oligosaccharide from the glycopeptide before MS.

The work presented here addresses glycoprotein

proteomics through multidimensional chromatography and MS of glycopeptides. Three types of problems are examined. One is the identification of parent proteins when the carbohydrate portion of glycopeptides is homogeneous. Another identifies proteins when the carbohydrate portion of the glycoconjugate is heterogeneous. And the third examines oligosaccharide heterogeneity at a particular site in a protein.

2. Materials and methods

2.1. Materials

Human serotransferrin, human serum, TPCK-treated trypsin, concanavalin A (Con A), *Bandeiraea simplicifolia* (BS-II) lectin, tris[hydroxymethyl]aminomethane (Tris base), tris[hydroxymethyl]aminomethanehydrochloride (Tris acid), iodoacetic acid, cysteine, dithiothreitol (DTT), *N*-tosyl-L-lysylchloromethyl ketone (TLCK), and *N*-acetyl-D-glucosamine were purchased from Sigma (St. Louis, MO, USA). LiChrospher Si 1000 (10 μm, 1000 Å) was obtained from E. Merck (Darmstadt, Germany). 3,5-Dimethoxy-4-hydroxy-cinnamic acid (sinipinic acid), 3-aminopropyltriethoxy silane, polyacrylic acid (PAA), and dicyclohexyl carbodiimide (DCC), and trideuteroacetic anhydride were purchased from Aldrich (Milwaukee, WI, USA). Methyl-α-D-mannopyranoside was obtained from Calbiochem (La Jolla, CA, USA). Toluene, dioxane and dimethyl sulfoxide (DMSO) were purchased from Fisher Scientific (Fair Lawn, NJ, USA). *N*-Hydroxyl succinimide (NHS) and trifluoroacetic acid (TFA) (HPLC grade) were purchased from Pierce (Rockford, IL, USA). HPLC-grade water and acetonitrile (ACN) were purchased from EM Science (Gibbstown, NJ, USA). All reagents used directly without further purification. Carboxypeptidase sequencing kits were obtained from PE-Biosystems (Framingham, MA, USA).

2.2. Synthesis of lectin columns

The procedure used here for lectin immobilized is a variant of other methods widely described in the literature for protein immobilization on silica-based supports [26]. A 1-g amount of LiChrospher Si 1000 was activated for 5 h at room temperature by adding 40 ml of 6 *M* HCl. The silica particles were then filtered and washed to neutrality with deionized water after which they were dried initially for 2 h at 105°C and then at 215°C overnight. Silica particles thus treated were reacted with 0.5% 3-aminopropyltriethoxy silane in 10 ml toluene for 24 h at 105°C to produce 3-aminopropyl silane-derivatized silica (APS silica).

PAA (0.503 g; M_r 450 000), *N*-hydroxysuccinimide (1.672 g), and DCC (6.0 g) were dissolved into 40 ml DMSO and shaken for 3 h at room temperature to derivatize the polymer with *N*-hydroxysuccinimide [27]. The reaction mixture was filtered and the activated polymer harvested in the supernatant. Acrylate polymer was then grafted to silica particles by adding the APS silica described above to the activated acrylate polymer. Following a 12 h reaction at room temperature, the particles were filtered and washed sequentially with 50 ml DMSO, 50 ml dioxane and 50 ml of deionized water. This procedure produces a polyacrylate-coated silica with residual *N*-acyloxysuccinimide groups, specified as NAS-PAA silica.

An immobilized Con A column was synthesized by adding NAS-PAA silica (0.5 g) to 10 ml of 0.1 *M* $NaHCO_3$ containing 0.2 *M* methyl-α-D-mannopyranoside and 200 mg Con A. The reaction was allowed to proceed with shaking for 12 h at room temperature, after which the immobilized Con A sorbent was recovered by centrifugation and was washed with 0.1 *M* (pH 7.5) Tris buffer. The sorbent was stored in 0.1 *M* (pH 7.5) Tris buffer with 0.2 *M* NaCl until it was used. The sorbent was packed into a stainless steel column (50 mm×4.6 mm) using the wash buffer and a high-pressure pump from Shandon Southern Instruments (Sewickley, PA, USA). The column was washed with 0.1 *M* Tris (pH 7.5) with 0.2 *M* NaCl containing 1 m*M* $CaCl_2$ and 1 m*M* $MnCl_2$.

An immobilized *Bandeiraea simplicifolia* (BS-II) lectin column was synthesized by adding NAS-PAA silica (0.3 g) to 10 ml 0.1 *M* $NaHCO_3$ containing 0.2 *M* *N*-acetyl-D-glucosamine and 20 mg BS-II lectin. The reaction was allowed to proceed with shaking for 12 h at room temperature, after which the immobilized lectin containing particles were isolated by centrifugation, washed with 0.1 *M* (pH 7.5) Tris buffer, and stored in 0.1 *M* (pH 7.5) Tris buffer with

0.2 *M* NaCl until it was used. The sorbent was packed into a stainless steel column (50 mm, 4.6 mm) using the wash buffer and a high-pressure pump from Shandon Southern Instruments. The column was washed with 0.1 *M* Tris (pH 7.5) with 0.2 *M* NaCl containing 1 m*M* $CaCl_2$ and 1 m*M* $MnCl_2$.

2.3. Proteolysis

Human serotransferrin (5 mg) and biological samples (50 μl) were reduced and alkylated in 1 ml of 0.2 *M* Tris buffer (pH 8.5) containing 8 *M* urea and 10 m*M* DTT. Larger amounts of biological extract were used in some cases when proteins were at lower concentration. After a 2 h incubation at 37°C, iodoacetic acid was added to a final concentration of 20 m*M* and incubated in darkness on ice for 2 more hour. Cysteine was then added to the reaction mixture to a final concentration of 40 m*M* and the reaction allowed to proceed at room temperature for 30 min. After dilution with 0.2 *M* Tris buffer to a final urea concentration of 3 *M*, TPCK-treated trypsin (2% of enzyme by mass to that of the protein) was added and incubated for 24 h at 37°C. Digestion was stopped by adding TLCK in a slight molar excess over that of trypsin.

2.4. Carboxypeptidase sequencing

Carboxypeptidase was reconstituted according to the kit manufacturer and separated into five different concentrations, with a 10-fold difference between them. The highest concentration of carboxypeptidase was 1 pmol/μl. A 1-μl volume of ''calibration 2'' peptide mixture was diluted into 5 μl by deionized water and a 0.5 μl aliquot was placed on the matrix-assisted laser desorption ionization (MALDI) plate at each spot.

2.5. Chromatography

All chromatographic steps were performed using an INTEGRAL Micro Analytical Workstation from PE Biosystems. Tryptic digested human serotransferrin (0.1 ml) was injected onto the Con A affinity column subsequent to equilibration with a loading buffer containing 1 m*M* $CaCl_2$, 1 m*M* $MgCl_2$, 0.2 *M* NaCl, and 0.1 *M* Tris–HCl (pH 7.5). The Con A column was eluted at 1 ml/min sequentially with two column volumes of loading buffer and then 0.2 *M* methyl-α-D-mannopyranoside in 0.1 *M* Tris (pH 6.0).

Analytes displaced from the affinity column with 0.2 *M* methyl-α-D-mannopyranoside [28] were directed to a 250×4.6 mm Peptide C_{18} analytical reversed-phase high-performance liquid chromatography (HPLC) column from PE Biosystems, which had been equilibrated for 5 min at 1.0 ml/min with 5% acetonitrile containing 0.1% aqueous TFA. The glycopeptides were then eluted at 1.0 ml/min in a 35 min linear gradient to 50% acetonitrile in 0.1% aqueous TFA. Eluted peptides were monitored at 220 nm and fractions manually collected for MALDI-MS analysis.

Tryptic digested human serum (0.2 ml) was injected on the Con A and reversed-phase HPLC column using conditions similar to those used with human serotransferrin with several exceptions. The reversed-phase column was washed for 10 min at 1 ml/min with 10% acetonitrile containing 0.1% aqueous TFA and the glycopeptides were eluted at 1 ml/min with a 120 min linear gradient to 70% acetonitrile containing 0.1% aqueous TFA.

Affinity selection with the BS-II lectin columns was achieved using 100 μl of nuclear extract. The column was pre-equilibrated with 0.2 *M* NaCl in 0.1 *M* Tris buffer (pH 7.5). After non-binding species had eluted to waste, glycans were eluted with 0.2 *M* *N*-acetylglucosamine. Direct transfer of the eluted glycan to a C_{18} Pepmap RPLC column was achieved by switching values. Glycopeptides were separated on the RPLC column using a 50 min gradient from 0.1% TFA with 1% acetonitrile to 0.1% TFA with 90% acetonitrile. The BS-II column was regenerated by washing with 20 ml starting buffer while the RPLC column was recycled with 5 ml of 0.1% TFA containing 1% acetonitrile.

2.6. Capillary electrophoresis

Glycopeptides from human serum were dissolved in 50 μl water and a fraction of the sample injected into a 50 cm fused-silica capillary operated at 15 kV with 10 m*M* phosphate buffer (pH 7.0). Each peak was collected and mixed with 1 μl of sinipinic acid [29] saturated matrix solution containing water–ace-

tonitrile (50:50) and 3% TFA. The mixture was spotted directly on a MALDI plate and analyzed by MALDI-MS.

2.7. Mass spectrometry

MALDI-MS was performed using a Voyager DE-RP BioSpectrometry Workstation from PE Biosystems. Samples were prepared by mixing a 1-μl aliquot with 1 μl of matrix solution. The matrix solution for glycopeptides was prepared by saturating a water–acetonitrile (50:50), 3% TFA solution with sinipinic acid. Samples of 1 μl were spotted into wells of the MALDI sample plate and allowed to air-dry before being placed in the mass spectrometer. All peptides were analyzed in the linear, positive ion mode by delayed extraction using an accelerating voltage of 20 kV, unless otherwise noted. External calibration was achieved using a standard ''calibration 2'' mixture from PE Biosystems. A 19-point data smoothing process was performed in many cases.

2.8. Deglycosylation by N glycosidase F (PGNase F)

Glycopeptides were freeze–dried overnight. A 10-μl volume of PGNase F was added to each tube and allowed to react during 12 h.

2.9. Partial C-terminal sequencing with carboxypeptidase (CBXP)

A carboxypeptidase sequencing kit from PE Biosystems was used in this work. When reconstituted, five different concentrations of CBXP differing 10-fold from each other were obtained. Aliquots of CBXP solutions were placed directly on samples in the MALDI plate wells. The highest concentration of carboxypeptidase used was 1 pmol/μl. A 1-μl volume of the ''calibration 2'' peptide mixture was diluted to 5 μl with deionized water. Samples of equal volume were placed in five wells on the MALDI plate. A 0.5-μl sample aliquot was put on the MALDI plate at each spot. After all samples were dry, 0.5 μl of the serial diluted carboxypeptidase solution was placed on the sample in each of the five wells. Peptide digestion was achieved directly on the MALDI plate [30]. Proteolysis stopped when the water evaporated from the spots, generally in 15 min. After drying, 0.5 μl α-cyano-4-hydroxy cinnamic acid saturated matrix solution was added to each well prior to mass spectral analysis.

2.10. Database searches

Database searches were executed with the Simulation Assisted, Online Protein Identification/quantification (SAOPI) software developed in our laboratory. This software is a portion of the Ph.D. Thesis of co-author X.Z. Actually, any software can be used that allows amino acids of unique molecular mass to be analyzed. In this case, the unique amino acids would be serine-GlcNAc and threonine-GlcNAc. Human protein databases were downloaded from Genbank using the Batch Entrez program at http://www.ncbi.nlm.nih.gov/Entrez/batch.htlm. Human proteins were virtually digested to generate tryptic peptides containing up to one missed cleavage site. Cysteine residues were assumed to be derivatize with iodoacetic acid. Mass accuracy was set a 1 u. In the case of searches for *N*-linked glycopeptides, peptides were selected having the sequence NXS/T where X is any amino acid residue except possible P or D.

3. Results and discussion

3.1. The analytical protocol

The general protocol chosen for these studies is illustrated in Fig. 1. Subsequent to disulfide reduction, alkylation, and proteolysis of all proteins in a sample, glycopeptide fragments were selected with an immobilized lectin column. Two types of lectin affinity columns were examined in this approach to proteomics. One was the case in which the lectin was of broad selectivity. Although Con A was used in this work, *Ricinus communis* lectin would have been of even broader selectivity. Con A has high affinity for *N*-type hybrid and high-mannose oligosaccharides, slightly lower affinity for complex diantennary oligosaccharides, and virtually no affinity for complex *N*-type tri- and tetraantenary-oligosaccharides [31]. It is ideal for selecting glycopeptides from digests of *N*-type glycoproteins. Most of the *N*-type

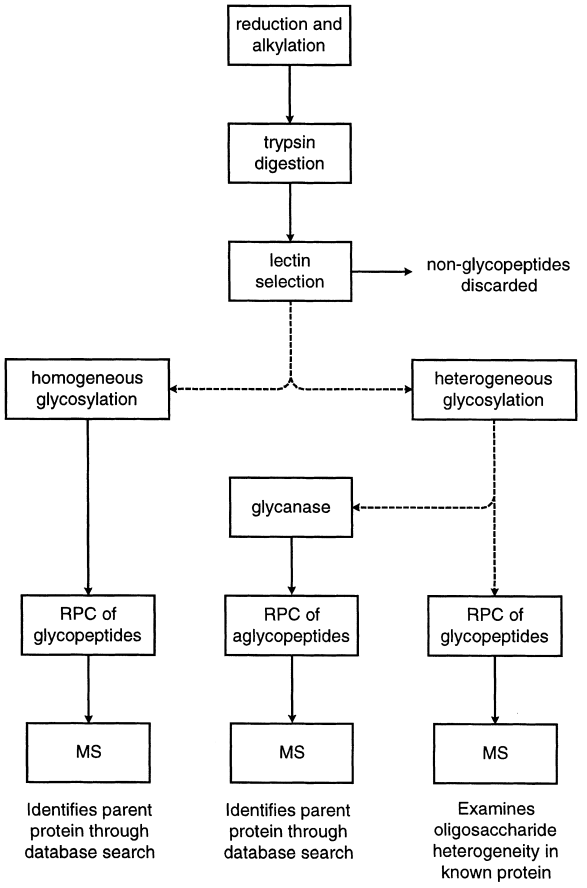

Fig. 1. The general analytical protocol.

glycoproteins contain glycoforms that are recognized by Con A. This means that any particular peptide will likely appear in multiple glycoforms. A broad selectivity lectin was used because it was the objective to select as many glycopeptides as possible.

The other type of immobilized lectin examined in these studies was of narrow selectivity, generally targeting a single type of oligosaccharide. The lectin from *Bandeiraea simplicifolia* (BS-II) was used in this situation. BS-II shows high selectivity for *N*-acetylglucosamine (GlcNAc) derivatized oligosaccharides [32]. When a single carbohydrate is selected, it is generally of know structure and mass, as with GlcNAc. Although the mixture selected may be very complex in terms of the peptide portion of the glycoconjugate, the carbohydrate will be identical in all cases.

The glycopeptide fraction selected from a trypsin digest of a complex protein mixture was always itself observed to be complex. This means that the glycopeptide fraction released from a lectin column would be too complex to be amenable to direct analysis by MS. All samples were further resolved by RPLC, either directly from the affinity column or after cleavage of the carbohydrate from the glycoconjugate. Fractions were collected from the RPLC column and analyzed by MALDI.

3.2. Identification of known glycoproteins

Serotransferrin, i.e., transferrin from human serum, was chosen as a model protein for the examination of glycopeptide selection. Human serotransferrin is a glycoprotein of M_r 80 000 containing 679 amino acids with *N*-glycosylation sites at asparagine 413 and asparagine 611. It is well known that this protein can be selected from a complex protein mixture with antibodies. One of the disadvantages of lectins is that they select a large number of glycoproteins that must be further fractionated to identify a specific protein. Because peptides are easier to resolve and more amenable to identification by MS than proteins, identification of serotransferrin through one of it's glycopeptides was examined.

Resolution of all the tryptic peptides of transferrin by RPLC is seen in Fig. 2a. Subsequent to chromatography on an immobilized Con A column, this relatively complex digest is reduced to a mixture of two components according to RPLC (Fig. 2b). It should also be noted that the peaks are slightly broad. Peptides glycosylated at residues asparagine 413 and asparagine 611 eluted from the RPLC column at 27.5% and 33.4% of solvent B, respectively, based on the ensuing interpretation of mass spectral data.

MALDI-MS of the two major components from Fig. 2b produced the spectra seen in Fig. 3a and b, respectively. Although the chromatographic peaks appear to be homogeneous, MALDI-MS indicates considerable heterogeneity within the two fractions. This is probably the reason for the broad peaks in the reversed-phase chromatogram. The hydrophobic peptide backbone of glycopeptides is responsible for adsorption to RPLC sorbents. Oligosaccharides play

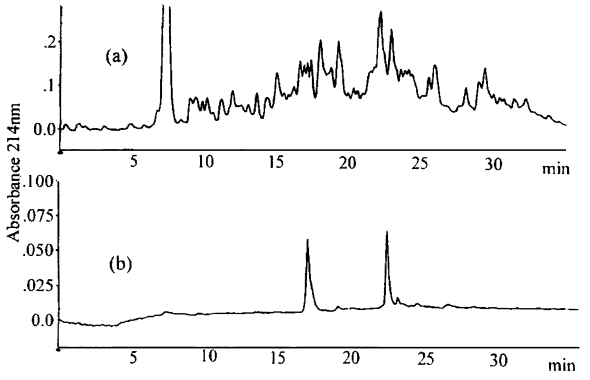

Fig. 2. Reversed-phase liquid chromatographic separation of peptides from human serotransferrin: (a) total tryptic digest, (b) glycopeptides selected from the tryptic digest with a Con A column. Glycopeptides were displaced from the Con A column with 0.2 *M* methyl-α-D-mannopyranoside and eluted directly into a 250×4.6 mm Peptide C_{18} analytical reversed-phase HPLC column. After sample deposition at the inlet of the reversed-phase column, the columns was equilibrated for 5 min at 1.0 ml/min with 5% acetonitrile containing 0.1% aqueous TFA. Peptides were then eluted at 1.0 ml/min in a 35-min linear gradient to 50% acetonitrile in 0.1% aqueous TFA. Eluted peptides were monitored at 220 nm and fractions manually collected for MALDI-MS analysis.

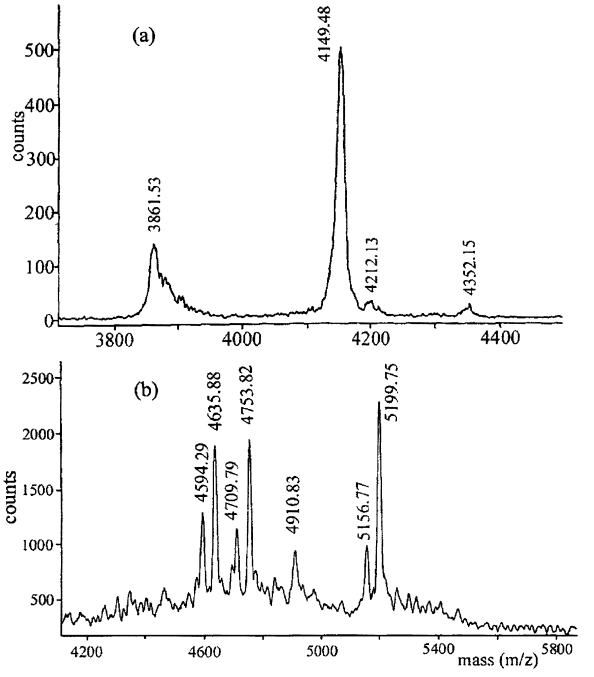

Fig. 3. Mass spectra of *N*-linked glycopeptides from human serotransferrin by a Con A lectin column: (a) the first glycopeptide in the chromatogram shown in Fig. 2b; (b) the second glycopeptide in Fig. 2b. Mass spectra were obtained as described in Materials and methods.

little role in retention except to make an overall contribution to analyte hydrophilicity, as will be shown below. Glycoforms are expected to be slightly resolved by RPLC with the more hydrophilic components eluting slightly ahead of less glycosylated forms. The great advantage of combining RPLC and MALDI-MS in the analysis of glycoforms is that RPLC discriminates between peptides whereas MALDI-MS sees heterogeneity within glycoforms.

Human blood serum is a complex protein mixture. Chromatograms in Fig. 4a and b show the substantial complexity of the glycopeptide mixture selected from a tryptic digest of human serum by a Con A affinity chromatography column [33]. Based on Fig. 2b, fractions eluting between 27 and 28% and 33 and 34% were collected from the reversed-phase column and their mass spectra compared to the human serotransferrin glycopeptides (Fig. 3a and b). Although extremely complex, mass spectra (Fig. 5a and b) obtained from fractions corresponding in chromatographic properties to the serotransferrin glycopeptides reveal the presence of these signature peptides in the serum sample.

Fig. 5a shows masses at 3861, 4153 and 4213 u, matching the glycopeptide peaks from Fig. 2. Peak intensities were lower and the resolution poorer in

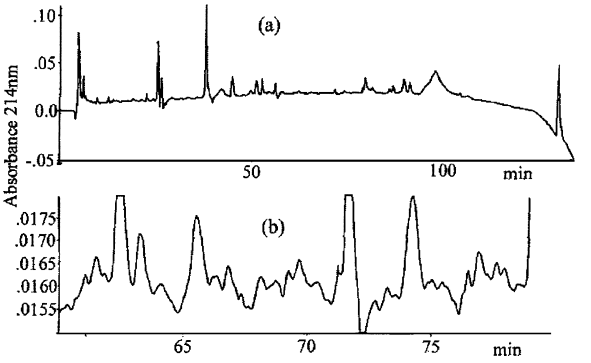

Fig. 4. Reversed-phase chromatograms of *N*-linked glycopeptides selected from human blood serum by a Con A lectin column: (a) total chromatogram, (b) expanded chromatogram. The selection process and chromatographic conditions are identical to those described in Fig. 2b.

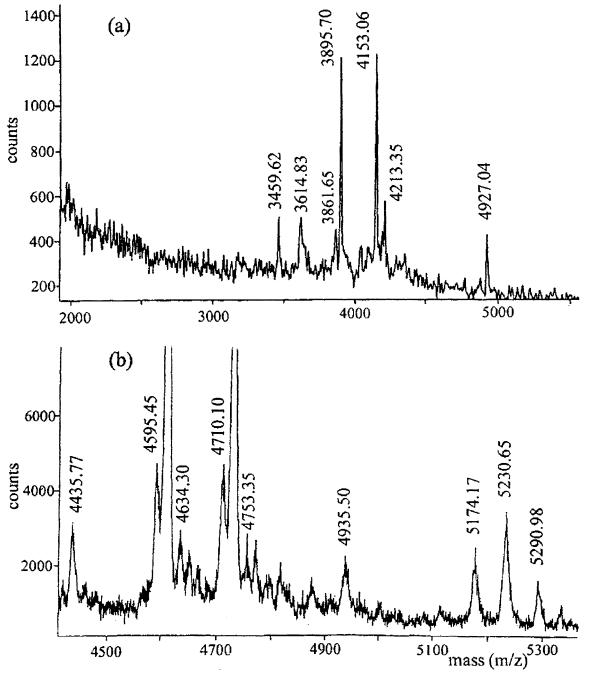

Fig. 5. Mass spectra of *N*-linked glycopeptides selected from human serum by Con A and separated by RPLC: (a) glycopeptides collected from RPLC between 27 and 28% of solvent B: (b) glycopeptides collected from 33 to 34% of solvent B. Mass spectra were obtained as described in Materials and methods.

the serum sample due to the relatively small amount of transferrin. Higher laser power was used to generate spectra than with pure human transferrin. Because laser power impacts relative peak intensities, it is not possible to make a direct quantitative comparison between these two samples. Spectra were smoothed by a 19-point averaging process to increase the signal-to-noise ratio. This caused the mass error to be a little higher. Glycoforms at 3459, 3614 and 3895 μ were either absent, or suppressed sufficiently to not be seen in the serum sample. If the absence of these peaks is due to variations between the protein standard and the serum sample, this is valuable analytical data. Variations due to differential quenching of glycoforms would be more difficult to solve.

It appears from this data that glycopeptides can be used as an analytical surrogate of the parent protein from which they were derived. This approach would be particularly well suited to the identification of proteins that are of limited solubility or difficult to purify. The limitation of this method is that it cannot be used directly for the identification of unknown proteins. When the mass of both the peptide and carbohydrate portions of the conjugate are unknown, database identification is precluded. Although databases contain polypeptide sequence, there is no information on glycan structure.

3.3. Analysis of glycoproteins with glycans of unknown structure

The problem of identifying glycopeptides of unknown glycosylation has been noted above. One approach to this problem is to remove the carbohydrate portion of the glycoconjugate. Hydrolysis of glycans from the peptide backbone has been achieved in two ways. Trifluoromethlysulfonic acid (TFMS) has been used widely with glycoproteins where it is reported to cleave oligosaccharides from proteins without disrupting the polypeptide backbone [34]. This technique was examined with glycopeptides in these studies and found to be unsatisfactory. A mixture of products were generated, among them partial cleavage fragments in the oligosaccharide portion of the conjugate and cleavage fragments within the peptide backbone (data not presented). Conditions were not found in which hydrolysis was limited to a single site.

Enzymatic digestion is the other approach to hydrolysis. In the case of *N*-glycosidation, PNGase F has been widely used to cleave *N*-linked glycoconjugates at asparagine; producing an oligosaccharide and a peptide with an aspartate residue [35]. Accessibility of the linkage site to PNGase F was not a problem, perhaps because the substrates were smaller glycopeptides. Glycans were easily cleaved from peptides. PNGase F efficacy was determined by RPLC and MS. Retention of the transferrin glycopeptides eluting at 27.4% and 34.4% acetonitrile concentration shift to 30.5% and 37.7%, respectively after deglycosylation. Peptides and glycopeptides were retained by RPLC columns whereas oligosaccharides were not. Peptide heterogeneity also decreased with deglycosylation as indicated by MS.

The second of the deglycosylated peptides from transferrin was selected for further characterization. After PNGase F digestion, the molecular mass of this peptide by MALDI was found to be 2964. This is

surprising in lieu of the fact that the sequence of the peptide was thought to be QQQHLFGSNVTDCSGNFCLFR (molecular mass= 2516). C-Terminal sequencing of this peptide with a commercial carboxypeptidase kit designed for MALDI-MS sequencing revealed a C-terminal sequence of H_2N-GYF-CO_2H. A search of the database matched this sequence and the mass of the parent peptide to the transferrin sequence QQQHLFGSNYTDCSGNF (molecular mass= 1939). The explanation for this unexpected finding was that the trypsin preparation had a small chymotrypsin impurity. During the 24 h incubation used in proteolysis, the peptide was also cleaved at phenylalanine. Similar findings were seen with the first peptide. These findings show the need for absolutely pure trypsin.

The utility of the approach outlined in Fig. 1 with a complex sample was examined using human serum. The sample was reduced, alkylated, and tryptic digested overnight. Glycopeptides selected by the Con A column were eluted directly onto an RPLC column. Ten percent acetonitrile with 1% TFA was used to wash sugars and salts from the RPLC column. Glycopeptides were then eluted from the RPLC column in a single fraction by switching to a mobile phase containing 60% ACN with 1% TFA in water, manually collected, and lyophilized. Deglycosylation was carried out by adding 25 μl PNGase F to the sample and incubating overnight. Peptides in the hydrolysate were separated on a reversed-phase column in a 100 min linear gradient ranging from 10% acetonitrile with 1% TFA to 60% acetonitrile with 1% TFA. Fractions were manually collected every minute and freeze–dried for mass spectral analysis.

Several fractions were randomly selected for further characterization. The fraction starting to elute at 37.2% of the B solvent was found by MALDI-MS (Fig. 6a) to have components with molecular masses of 861.3, 1460.9, 1564.7 and 2089.5, with the 1460.9 peak being dominant. Treatment of this sample on the MALDI plate with the carboxypeptidase sequencing enzymes produced a new set of peptides of molecular masses 1089.6, 1160.7, 1261.7 and 1332.6 (Fig. 6b). It is important to note again PNGase F cleavage converts an asparagine residue at the point of glycosylation to an aspartate residue [35]. By sequential subtraction it is seen that 1460.9−1332.6=128.3 (K), 1332.6−1261.7=70.9 (A), 1261.7–1160.7=101 (T), 1160.7−1089.6=71.1 (A). Thus, the C-terminal sequence of the dominant 1460.9 component is –ATAK. Using this data and assuming that (1) the peptide had the mandatory NXS/T sequence of an *N*-linked glycopeptide, (2) mass had been determined with an accuracy of no better than 2 μ, and (3) an asparagine residue at the point of glycosylation is converted to an aspartate residue, the most likely candidate in the database (swissprot:htp2_human) was NLFLNHSENATAK, a tryptic fragment of haptoglobin that is glycosylated on both asparagines residues in the peptide.

The fraction that started eluting from the RPLC column at 41.2% of the B solvent was also chosen for analysis. The major component in this fraction was of molecular mass 1796.3 according to MALDI-MS (Fig. 7a). Carboxypeptidase digestion produced a new set of peptides at molecular masses of 1442.1, 1272.1, 1157.1 and 1058.0 (Fig. 7b). Sequential subtraction revealed the partial sequence at the C-terminus to be VDIG/GI/LG/GL. Again it was assumed that (1) the peptide had the mandatory NXS/T sequence of an *N*-linked glycopeptide, (2) mass had been determined with an accuracy of no better than 2 μ, and (3) an asparagine residue at the point of glycosylation is converted to an aspartate residue. Using this partial sequence a search of the database found a match in the peptide VVL-HPNYSQVDIGLIK from haptoglobin.

Haptoglobin is composed of four disulfide linked subunits and exists in three forms; $\alpha_1\alpha_1\beta\beta$, $\alpha_1\alpha_2\beta\beta$, and $\alpha_2\alpha_2\beta\beta$. The α_1 and α_2 subunits are both M_r 9100 and β subunit is M_r 40 000. Oligosaccharides are reported to be attached exclusively to the β chain at positions 23, 46, 50 and 80 [36]. The glycan structures in haptoglobin are heterogeneous and can vary with disease state. This can impact the affinity and degree to which haptoglobin binds to lectins. Obviously the haptoglobin derived peptides identified above both arose from the β chain. Since the β subunit is common to all forms of haptoglobin, these peptides may not be used to differentiate between the phenotypes. Furthermore, the degree to which they are captured by Con A could vary in the cases of inflammatory diseases, liver diseases, and some types of cancer. This could actually be valuable as a

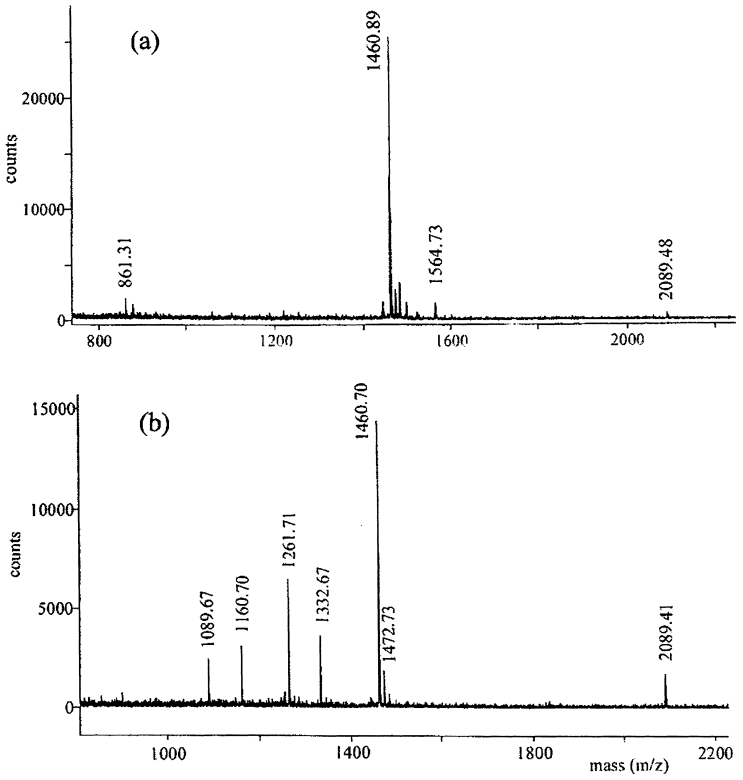

Fig. 6. Mass spectra of deglycosylated glycopeptides from serum collected from the RPLC column in the fraction eluting between 36 and 38% of solvent B: (a) before carboxypeptidase treatment, (b) after carboxypeptidase treatment. Carboxypeptidase treatment of samples on the MALDI plate, sample preparation for mass spectrometry, and conditions for operation of the mass spectrometer were as described in Materials and methods.

diagnostic tool. The procedure described above for selection of glycoconjugate peptides along with differential isotopic labeling for quantification [33] could be used to show that the amount of peptide captured by Con A is diagnostic of various diseases.

3.4. Analysis of glycoproteins with homogeneous glycosylation

The simplest case of homogeneous glycosylation is the one in which a single *N*-acetylglucosamine (GlcNAc) is coupled to a serine or threonine residue in the protein. A number of nuclear and cytoplasmic proteins are modified in this way, including nuclear-pore proteins, heat-shock protein, p53 tumor-suppressor, nuclear oncogene proteins, RNA polymerase II, and many transcription factors [37]. These proteins are thought to play a regulatory of role, particularly in nuclei [38].

Nuclear proteins with *O*-GlcNAc glycosylation were chosen for study. Nuclei were isolated from human U937 cells by centrifugation and the histones selectively removed by manipulation of the ionic strength. Proteins terminally glycosylated with GlcNAc were isolated according to the general protocol outlined in Fig. 1. Based on the fact that a lectin (BS-II) from *Bandeiraea simplicifolia* binds proteins and peptides with a terminal GlcNAc residue, a BS-II lectin column was used to select this type of glycan. Although other glycoproteins in nuclear extracts contain fucose, mannose, galactose, and glucose, *O*-GlcNAc derivatized proteins are virtually the only terminal GlcNAc containing species in either the cytoplasmic or nuclear compartments of cells [39]. After direct transfer from the

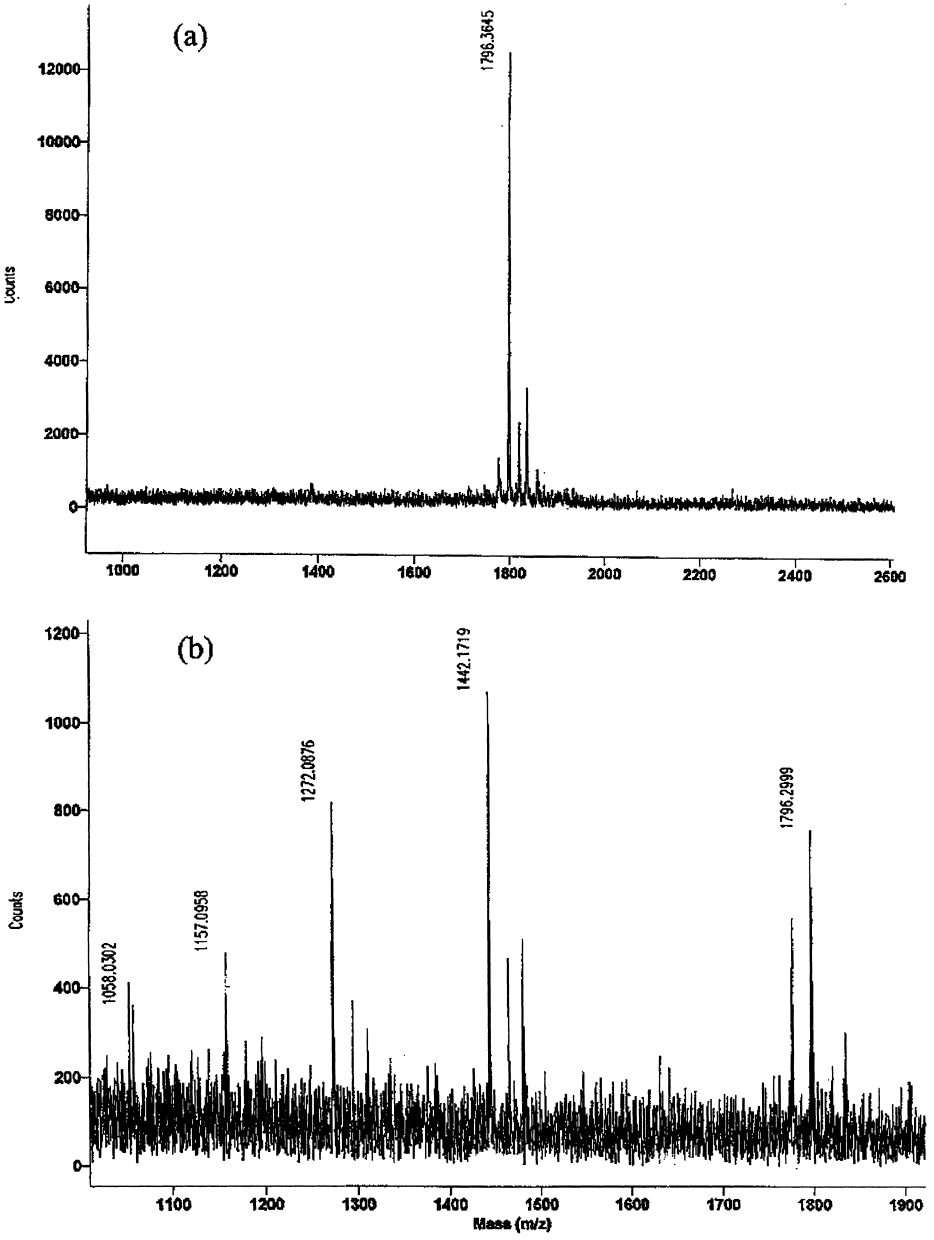

Fig. 7. Mass spectra of deglycosylated glycopeptides from serum collected from the RPLC column in the fraction eluting between 40 and 42% of solvent B: (a) before carboxypeptidase; (b) after applying carboxpeptidase. Experimental conditions were identical to those in Fig. 6.

affinity column to a reversed-phase column, the mixture of glycopeptides was further fractionated by gradient elution and fractions collected for MALDI-MS. The great advantage of this affinity chromatography approach is that it allows both fractionation and concentration of glycopeptides. This is particularly useful in the case of nuclear extracts where the concentration of these species is low.

One of the fractions collected from the RPLC column was examined by MS. The MALDI-MS

spectrum of the fraction eluting at 26.7% acetonitrile showed four major peaks with molecular masses of 2721.2, 2569.9, 2456.3, and 1451.9, respectively (Fig. 8a). Because *O*-glycosylation with GlcNAc involves a single carbohydrate residue, there are neither glycoforms nor mass heterogeneity within the glycan portion of the molecule. Each peak in the MALDI mass spectrum is from a different peptide, but multiples of 203 μ for the mass of the GlcNAc residue must be subtracted from the mass of peptides as part of the database search routine. This is because proteins may be glycosylated with GlcNAc at multiple sites and databases contain only primary structure information. In order to deal with this situation, mass 204 was subtracted for every serine (S) or threonine (T) in a peptide candidate and all possible combinations beyond one GlcNAc are considered. Peptides also sometimes picked up an H^+, Na^+, and K^+ ions. This problem is complicated by the fact that acquisition H^+, Na^+, or K^+ ions can even vary across a single MALDI well [40] (data not shown). On one side of the well a peptide may have acquired an H+ ion while peaks on the other side of the well are due to acquisition of Na+ and K+ ions. These variables must also be considered in database search routines.

Carboxypeptidase digestion of peptides spotted on MALDI plates was used to partially sequence the peptides noted above. In this process, the peptide mixture is placed in several wells on the plate and treated with increasingly diluted concentrations of a mixture of carboxypeptidases [30]. Multiple dilutions are used to ensure that one will be of proper concentration to digest peptides on the plate. This mixture of enzymes digests peptides from the C-terminus, producing a mixture of increasingly truncated peptide products varying by one amino acid.

It was observed (Fig. 8b) that the parent ions of

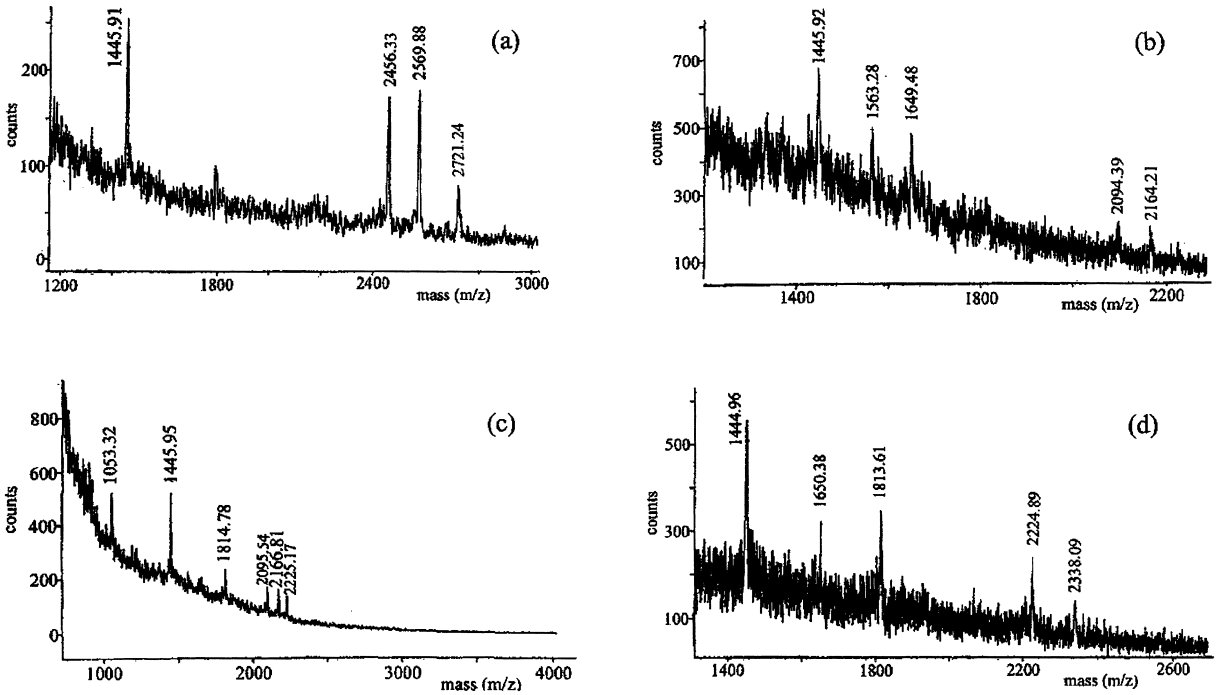

Fig. 8. Mass spectra of *O*-linked glycopeptides from a nuclear extract that were selected from with a BS-II lectin column and collected from an RPLC column in the fraction eluting between 26 and 28% of solvent B: (a) before applying carboxypeptidase; (b) after applying carboxypeptidase at low concentration; (c) after applying carboxypeptidase at a higher concentration; (d) after applying carboxypeptidase at a still higher concentration. The chromatographic protocol and elution conditions were identical to those described in Fig. 2. Concentrations of carboxypeptidase and the digestion protocol are described in Materials and methods. Application of the MALDI matrix and conditions for operation of the mass spectrometer are described in Materials and methods.

molecular masses of 2456.3, 2569.9 and 2721.2 disappeared after carboxypeptidase treatment, while the parent ion with molecular mass of 1451.9 remained. Because all the truncated products developed were of mass higher than 1451.9, it was concluded that they were derived from the parent ions 2456.3, 2569.9 and 2721.2, respectively. There were two clusters of carboxypeptidase-cleaved products. One cluster of masses was at 1563.3, 1649.5 and 1814.8 and the other at 2094.4, 2164.2, 2225.2 and 2338.1. Since the truncated products ranged in molecular mass from 1563.2 to 2338.9, they could have been generated from the parent ions 2456.3, 2569.9 and 2721.2, respectively. There were two clusters of carboxypeptidase cleaved products. One cluster is at masses 1563.3, 1649.5 and 1814.8 and the other at 2094.4, 2164.2, 2225.2 and 2338.1. Because it is highly unlikely that the products from either cluster could have been derived from two different parent peptides, it was concluded that each cluster came from a different peptide.

Peptides in the second cluster produced by carboxypeptidase cleavage (Fig. 8b–d) permitted a partial sequence to be established; 2338.1−2225.2=113 (I/L/N), 2225.2−2164.2=61 (G) and 2164.2−2094.4=70 (A). This suggests a partial sequence of AG (I/L/N) derived from a tryptic peptide of either 2456.3, 2569.9 or 2721.2 molecular mass with at least one T or S and a C-terminus of either K or R. The partial sequence –TW– could also be part of this peptide, but does not have to be. (In the case of chymotrypsin contaminated trypsin, F could also be at the C-terminus). Because GlcNAc was not found in any of the carboxypeptidase fragments, the parent peptide also had to have at least one –S– or –T– residue with an attached GlcNAc. Considering that (i) the mass of one or more GlcNAc residues had to be subtracted, (ii) the possibility of H^+, Na^+ and K^+ acquisition during the MALDI process, (iii) mass error might be up to 3 μ, and (iv) the peptide would have to have ladder fragments of 2094.4, 2164.2, 2225.2 and 2338.1 μ produced by carboxypeptidase, tryptic peptides from the 70 000 proteins in the human database were matched to the experimentally derived data. The peptide QSPPESEILVHCSAGIGR, after addition of three GlcNAc residues and one Na^+ ion, was the only one found in the database that matched the experimental data. The peptide reported in the database spanned positions 105 to 122 of a protein-tyrosine phosphatase (Acc: 300036) taken from a human leukemia cell line F-36 P. (The nuclear extract used in these experiments was also derived from a human cancer cell line). The three serine residues in the peptide correspond to positions 106, 110, and 117 in the parent protein. Based on the fact that carboxypeptidase did not digest past the alanine residue at position 118, it is likely that the GlcNAc on serine 117 either reduced or block cleavage by carboxypeptidase. This could be useful in locating glycosylation position in glycopeptides.

Differences between adjacent peaks in the first cluster (Fig. 8b) allowed several amino acids to be identified; 1814.8−1649.5=165.3 (W) and 1649.5−1563.3=86 (T). Based on the fact that these peptides could not have come from the parent peptides with masses of 2569.9 and 1451.9 in Fig. 8a, they must have been derived from either the 2456.3 or 2721.2 molecular mass peptides. Database searches produced no peptides that matched the experimental data. It must be concluded that the protein from which the parent peptide and carboxypeptidase fragments were derived is not in any of the known databases.

4. Conclusions

Selection in affinity chromatography generally targets particular molecular structures. The results presented here show this is an advantage in proteomics when the objective is to identify proteins with specific types of post-translational modification.

Selection of glycopeptides from tryptic digests of complex protein mixtures by lectin-based affinity chromatography columns can be an effective route to glycoprotein proteomics when coupled with RPLC and MS. This is especially true when a small amount of additional information about the peptide portion of the glycoconjugate can be gathered from either the chromatographic behavior of standards or a partial sequence can be derived from MS. The advantage of the technique is that it is fast and targets specific classes of glyoptroteins. Limitations of the method are that it does not discriminate between glycoforms of proteins and that only the particular type of post-translational modification selected can be iden-

tified. It remains to be determined whether affinity selection will be of utility in cases where different post-translational modifications, such as phosphorylation and acylation, reside within the same peptide.

Acknowledgements

The authors greatly acknowledge financial support from the NIH grant GM-59996 and PE Biosystems.

References

[1] M.J. Page, B. Amess, C. Rohlff, C. Stubberfield, R. Parekh, Drug Discovery Today 4 (1999) 55.
[2] W.P. Blackstock, M.P. Weir, Trends Biotechnol. 17 (1999) 121.
[3] M. Quadroni, P. James, Electrophoresis 20 (1999) 664.
[4] J.R. Yates III, Trends Genet. 16 (2000) 5.
[5] J.E. Celis, Seibutsu Butsuri Kagaku 43 (1999) 213.
[6] R.A. Van Bogelen, E.E. Schiller, J.D. Thomas, F.C. Neidhardt, Electrophoresis 20 (1999) 2149.
[7] M. Geng, J. Ji, F.E. Regnier, J. Chromatogr. A 870 (2000) 295.
[8] A.A. Gooley, N.H. Packer, Proteome Res. (1997) 65.
[9] N.H. Packer, A. Pawlak, W.C. Kett, A.A. Gooley, J.W. Redmond, K.L. Williams, Electrophoresis 18 (1997) 452.
[10] Z. Keresztessy, J. Hughes, L. Kiss, M.A. Hughes, Biochem. J. 314 (1996) 41.
[11] D.M. Alperin, H. Latter, H. Lis, N. Sharon, Biochem. J. 285 (1992) 1.
[12] H. Matsue, K. Takagaki, K. Honda, Y. Nakagawa, F. Gejyo, M. Arakawa, M. Endo, Clin. Chem. (Winston-Salem, NC) 33 (1987) 2214.
[13] K. Yamamoto, T. Tsuji, T. Osawa, Mol. Biotechnol. 3 (1995) 25.
[14] T. Watanabe, N. Kondo, K. Kano, Biol. Plant 26 (1984) 99.
[15] S. Iturbe, S. Narasimhan, J.M. Merrick, J.A. Falk, M. Letarte, J. Immunol. 136 (1986) 4588.
[16] B.J. Harmon, X. Gu, D.I.C. Wang, Anal. Chem. 68 (1996) 1465.
[17] Z. El Rassi, Electrophoresis 20 (1999) 3134.
[18] K.F. Greve, D.E. Hughes, B.L. Karger, J. Chromatogr. A 749 (1996) 237.
[19] M.G. O'Shea, M.K. Morell, Chem. Aust. 63 (1996) 342.
[20] M. Jaquinod, T. Las Holtet, M. Etzerodt, I. Clemmensen, H.C. Thogersen, P. Roepstorff, Biol. Chem. 380 (1999) 1307.
[21] J. Colangelo, R. Orlando, Anal. Chem. 71 (1999) 1479.
[22] F.E. Regnier, A. Amini, A. Chakraborty, M. Geng, J. Ji, L. Riggs, C. Sioma, S. Wang, X. Zhang, LC·GC Mag. (2000) submitted for publication.
[23] C. Schaumann, F. Oesch, K.K. Unger, R.J. Wieser, J. Chromatogr. 646 (1993) 227.
[24] R. Apweiler, H. Hermjakob, N. Sharon, Biochim. Biophys. Acta 1473 (1999) 4.
[25] M. Geng, J. Ji, F.E. Regnier, J. Chromatogr. A 870 (2000) 295.
[26] A.L. Crumbliss, J. Stonehuerner, R.W. Henkens, J.P. O'Daly, J. Zhao, New J. Chem. 18 (1994) 327.
[27] L. Tan, Ph.D. Thesis, Purdue University, Lafayette, IN, 1998, p. 78.
[28] R.A. Merz, M. Horsch, H.P. Ruffner, D.M. Rast, Phytochemistry 52 (1999) 211.
[29] O. Vorm, P. Roepstorff, M. Mann, Anal. Chem. 66 (1994) 3281.
[30] D.H. Patterson, G.E. Tarr, F.E. Regnier, S.A. Martin, Anal. Chem. 67 (1995) 3971.
[31] M. Stratford, C.J. Bond, Biotechnol. Bioeng. 40 (1992) 835.
[32] H. De Boeck, F.G. Loontiens, F.M. Delmotte, C.K. De Bruyne, FEBS Lett. 126 (1981) 227.
[33] M. Geng, J. Ji, F.E. Regnier, J. Chromatogr. A 870 (2000) 295.
[34] A.S.B. Edge, C.R. Faltynek, L. Hof, L.E. Reichert Jr., P. Weber, Anal. Biochem. 118 (1981) 131.
[35] S. Hoffstetter-Kuhn, G. Alt, R. Kuhn, Electrophoresis 17 (1996) 418.
[36] G.A. Turner, Adv. Exp. Med. Biol. 376 (1995) 231.
[37] A. Varki, R. Cummings, J. Esko, H. Freeze, G. Hart, J. Marth (Eds.), Essentials of Glycobiology, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1999, Chapter 14.
[38] F.I. Comer, G.W. Hart, Biochim. Biophys. Acta 1473 (1999) 161.
[39] G.W. Hart, K.D. Greis, Pure Appl. Chem. 67 (1995) 1637.
[40] M. Geng, Ph.D. Thesis, Purdue University, Lafayette, IN, 2000, p. 131.