

A New Strategy for Identification of N-Glycosylated Proteins and Unambiguous Assignment of Their Glycosylation Sites Using HILIC Enrichment and Partial Deglycosylation

Per Hägglund,[†] Jakob Bunkenborg,[‡] Felix Elortza,[†] Ole Nørregaard Jensen,[†] and Peter Roepstorff^{*,†}

Department of Biochemistry and Molecular Biology; University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark, Present address: Institute of Genetic Medicine, Johns Hopkins University, 600 N. Wolfe St., Baltimore, MD-21287

Received November 26, 2003

Characterization of glycoproteins using mass spectrometry ranges from determination of carbohydrate-protein linkages to the full characterization of all glycan structures attached to each glycosylation site. In a novel approach to identify N-glycosylation sites in complex biological samples, we performed an enrichment of glycosylated peptides through hydrophilic interaction liquid chromatography (HILIC) followed by partial deglycosylation using a combination of endo- β -N-acetylglucosaminidases (EC 3.2.1.96). After hydrolysis with these enzymes, a single N-acetylglucosamine (GlcNAc) residue remains linked to the asparagine residue. The removal of the major part of the glycan simplifies the MS/MS fragment ion spectra of glycopeptides, while the remaining GlcNAc residue enables unambiguous assignment of the glycosylation site together with the amino acid sequence. We first tested our approach on a mixture of known glycoproteins, and subsequently the method was applied to samples of human plasma obtained by lectin chromatography followed by 1D gel-electrophoresis for determination of 62 glycosylation sites in 37 glycoproteins.

Keywords: proteomics • post-translational modifications • mass spectrometry • HILIC • endoglycosidase • lectin affinity chromatography • glycosylation • plasma proteins

Introduction

Genome sequencing has provided a blueprint onto which peptide sequence data and information about post-translational modifications must be correlated. The covalent attachment of carbohydrate chains is a ubiquitous co- and post-translational protein modification which frequently modulate the structures and functions of proteins and also infer protein-carbohydrate interactions.¹ Notably, changes in glycosylation are known to affect the immune system and altered glycosylation is implicated in several inflammatory diseases, cancers, and congenital disorders.^{2–4}

Unlike nucleic acids and proteins, carbohydrates are synthesized without an underlying template but guided by the combination of glycosyltransferases and glycosidases present in the cell at any particular time. Accordingly, glycosylation is a highly complex and dynamic process. Oligosaccharide chains are covalently attached to serine or threonine residues (O-glycosylation), to tryptophan residues (C-glycosylation) or to asparagine residues (N-glycosylation). N-glycosylation only takes place at asparagines appearing in the consensus-sequence NX-S/T/C where X can be any amino acid except proline.^{5,6} In a survey of the SWISS-PROT database by Apweiler,⁷ two-thirds

of all entries were found to contain at least one N-glycosylation sequon with an expected site occupancy rate of 2/3. This suggests that at least 50% of all proteins are glycosylated. Nevertheless, the number of proteins annotated as glycoproteins is only ~10%. This clearly illustrates the need for improved methods in glycosylation site mapping.

A typical asparagine-linked carbohydrate usually contains 7–25 monosaccharide units, 5 of which constitute the conserved trimannose-chitobiose core common to most N-glycans.⁸ There is a large degree of heterogeneity in the glycan structure and N-linked glycans are generally classified as high mannose, hybrid and complex type. The high mannose type is composed of one or several mannose residues attached to the core structure. Complex type N-glycans may contain several galactose, N-acetylglucosamine (GlcNAc), sialic acid and fucose residues attached to the core structure. Hybrid type glycans share similarities with both high mannose and complex type glycans.⁸ Frequently, an α -1–6 linked fucose residue is also attached to the innermost GlcNAc in complex and hybrid type glycans.

The full characterization of all glycoproteins in a complex sample is a very difficult task involving the determination of glycan size, branching points and linkages of the carbohydrate monomers for each glycosylation site. The current procedure for a full characterization requires purification of the protein to purity, followed by release of the glycans by either chemical

* To whom correspondence should be addressed. E-mail: roe@bmb.sdu.dk.

[†] Department of Biochemistry and Molecular Biology.

[‡] Institute of Genetic Medicine.

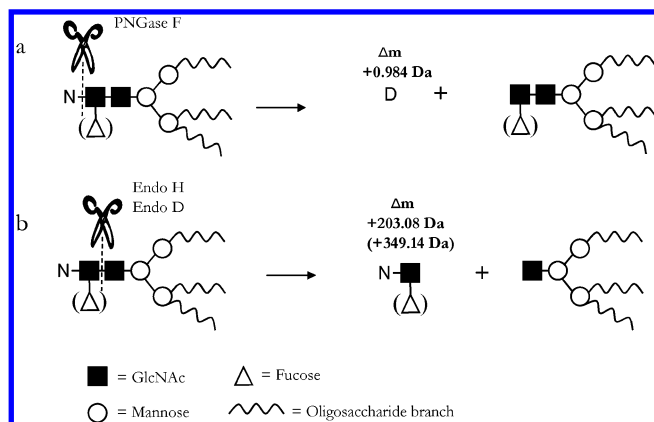


Figure 1. Enzymatic digestion of glycopeptides by endoglycosidases. This schematic picture only shows the conserved core structure of N-glycosylation with a putative fucose residue attached to the innermost GlcNAc residue. (a) Peptide-*N*-glycosidases, like PNGase F, hydrolyze the bond between asparagine and the innermost GlcNAc residue. In the reaction the asparagine residue is converted to aspartate, yielding a mass increase of 0.98 Da. (b) Endo- β -*N*-acetylglucosaminidases, like Endo D and Endo H, cleave the bond between the two GlcNAc residues in the core structure. Thus a single GlcNAc residue (or a fucosylated GlcNAc) is retained on the peptide, providing a mass increment of 203.08 Da (or 349.14 Da in the case of fucosylated structures).

or enzymatic cleavage. The pool of glycans are then often analyzed either by high performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD) or by matrix-assisted laser desorption/ionization mass spectrometry (MALDI-TOF MS) after treatment with an array of different exo-glycosidases.^{9–11}

One strategy for characterization of glycosylation sites in complex protein mixtures is to use some type of glycoprotein/glycopeptide enrichment method (e.g., lectin chromatography) combined with endoglycosidase digestion.^{12,13} Another approach is to use precursor ion scanning methods based on detection of glycan-specific fragment ions,^{14,15} which also has been used in combination with glycosidase digestion.¹⁶ However, the specificity of glycan-specific precursor ion scanning may be limited due to the presence of interfering ions.¹⁵ In this study, we have developed a novel approach for determination of glycosylation sites in complex samples. The method is essentially based on two steps: 1. Glycopeptide enrichment using hydrophilic interaction liquid chromatography (HILIC). 2. Enzymatic deglycosylation with a combination of endo- β -*N*-acetylglucosaminidases.

HILIC has previously been used mainly as a tool for analysis of small, polar molecules (e.g., carbohydrates, oligonucleotides, and amino acids), but also for separation of peptides and glycopeptides.^{17–21} Here, we have explored the use of HILIC to reduce the complexity of peptide/glycopeptide mixtures through depletion of hydrophobic peptides, and retention of hydrophilic glycopeptides. We have packed the HILIC material into GE-Loader tips since such microcolumns previously has proven very efficient for improving sensitivity and hence sequence coverage for proteins separated by gel-electrophoresis.²²

Enzymatic release of N-linked glycans with peptide *N*-glycosidase F (PNGase F; EC 3.5.1.52) is commonly used for determination of glycosylation sites.^{12,13,23–25} In the enzymatic reaction, the asparagine residue is converted to an aspartate residue with a concomitant mass increase of 0.98 Da (Figure

1a). One of the pitfalls of this approach is that it does not distinguish between enzymatic induced asparagine-to-aspartate conversion and other causes of deamidation, either in vivo or in vitro. In this study, we have instead performed enzymatic deglycosylation using endo- β -*N*-acetylglucosaminidases, which cleave the glycosidic bond between the two proximal GlcNAc residues in the chitobiose core, and leave one GlcNAc residue (and in some cases, a fucosylated GlcNAc) attached to the peptide (Figure 1b). This type of enzyme has previously been used in conjunction with mass spectrometric analysis of glycopeptides.²⁶ One of the advantages with endo- β -*N*-acetylglucosaminidase, is that the remaining GlcNAc residue can be used as a diagnostic marker of glycopeptides. The oxonium ion of GlcNAc at m/z 204.08 can be used for precursor ion scanning or to indicate the presence of an N-glycan attachment site in a peptide, i.e., by manual inspection of fragment ion spectra after database searches. Furthermore, the GlcNAc residue can be used as a variable modification in protein database search software (e.g., Mascot), or as a platform to perform various derivatization reactions. To test the potential of our approach we have analyzed proteins from human plasma enriched in glycoproteins by lectin chromatography. A total of 62 glycosylation sites were unambiguously identified among 37 glycoproteins.

Experimental Section

Proteins. Bovine ribonuclease B (RNase B), human α -1-acid glycoprotein and chicken ovalbumin were obtained from Sigma (St Louis, MI). Bovine fetuin was from Roche (Mannheim, Germany). Human plasma was obtained from a healthy donor.

In-Solution Digestion. 1 nmole of each protein was resuspended in 21 μ L 400 mM NH_4HCO_3 (pH 7.8) in 8 M urea. 5 μ L DTT (45 mM) was added and the mixture was incubated for 15 min at 50 °C, chilled, and 5 μ L of iodoacetamide (100 mM) was added, followed by an incubation in the dark at RT for 15 min. 140 μ L of H_2O and 20 pmole of sequence grade trypsin (Promega, Madison, WI) in 5 μ L was added and incubated overnight at 37 °C. The digests were then kept at –20 °C. Aliquots of 10 pmole were used for experiments.

Lectin Chromatography. 50 μ L agarose-bound lectin concanavalin A (Con A) isolated from *Canavalia ensiformis* (Vector Laboratories, Burlingame, CA) was washed three times in binding buffer (50 mM Hepes pH 7.5, 1 mM MgCl_2 , 1 mM CaCl_2 , 100 mM NaCl). After this, the supernatant was removed, 300 μ L human plasma and 600 μ L binding buffer was added, and the sample was incubated at 4 °C for 6 h. The sample was washed three times in binding buffer and finally elution of bound proteins was carried out with SDS-PAGE gel loading buffer (125 mM Tris-HCl (pH 6.8), 0.1% (w/v) bromophenol blue, 10% (v/v) 2-mercaptoethanol, 20% (v/v) glycerol, 4% (w/v) SDS).

SDS-PAGE and In-Gel Digestion. Proteins eluted from the lectin resin were separated by SDS-PAGE (12%) and stained with silver nitrate.²⁷ Protein gel pieces were cut out and placed in Eppendorf tubes. 30 μ L acetonitrile (ACN) was added and incubated for 10 min followed by removal of the supernatants. 50 μ L 10 mM DTT (in 100 mM NH_4HCO_3 (pH 7.8)) was added, incubated for 45 min at 56 °C, and then chilled to RT. The supernatant was removed and 30 μ L of freshly prepared 55 mM iodoacetamide in 100 mM NH_4HCO_3 (pH 7.8) was added and incubated in the dark for 30 min (RT). The supernatant was removed and 30 μ L of ACN was added. The supernatant was again removed and the sample was put in a speed-vac for 10

min. The gel pieces were rehydrated in chilled 100 mM NH_4HCO_3 (pH 7.8) (just enough to cover the gel-pieces) containing 12.5 ng/ μL sequence grade trypsin (Promega) and kept on ice for 45 min. The supernatant was removed and 100 mM NH_4HCO_3 pH (7.8) was added (just enough to cover the gel-pieces). The tryptic digestion was carried out at 37 °C overnight and then stored at −20 °C.

After the tryptic digestion the supernatant was removed and transferred to a new tube. 100 mM NH_4HCO_3 (pH 7.8) was added to cover the gel-pieces and incubated for 10 min. Then the same volume of ACN was added and incubated for 10 min. After incubation, the supernatant was collected and the gel-pieces were covered with 5% formic acid (FA) for 10 min before the same volume of ACN was added. After incubation for 10 min the supernatant was collected. The pooled supernatants were then lyophilized in a speed-vac and stored at −20 °C.

Microcolumn Chromatography. ZIC-HILIC chromatography media was a gift from Sequant (Umeå, Sweden). HILIC microcolumns were prepared by packing ZIC-HILIC media (particle size 10 μm) into GELoader tips (Eppendorf, Hamburg, Germany) in analogy to the methods used for reversed phase media and graphite powder.^{22,28} In-solution and in-gel digested samples were redissolved in 80% ACN, 0.5% FA and loaded onto HILIC microcolumns equilibrated with 80% ACN, 0.5% FA. The column was washed three times in 80% ACN, 0.5% FA and the bound peptides were eluted in 99.5% H_2O , 0.5% FA. Microcolumns packed with R2 and R3 reversed phase chromatographic media (Applied Biosystems, Foster City, CA) were equilibrated in 5% FA and samples were eluted in 70% ACN, 5% FA.

Endoglycosidase Digestion. All glycosidase digestions with endo- β -N-acetylglucosaminidases were carried out using the following mixture of enzymes (referred to as Endo D/H): 1 mU endoglycosidase D from *Streptococcus pneumonia* (ICN Biomedicals, Aurora, Ohio); 25 mU endoglycosidase H from *Streptomyces plicatus* (Roche); 2.5 mU neuraminidase from *Arthrobacter ureafaciens* (Roche); 0.5 mU β -galactosidase from *Diplococcus pneumonia* (Roche); 0.5 mU N-acetyl- β -glucosaminidase from *D. pneumonia* (Roche). The digestions were carried out at 37 °C in 100 mM NH_4Ac buffer (pH 5.5) overnight. Incubations with 1 mU PNGase F from *Flavobacterium meningosepticum* (Roche) were carried out at 37 °C in 100 mM NH_4CO_3 buffer (pH 7) overnight, with 1 mM phenyl methyl sulfonyl fluoride (Roche) added as a protease inhibitor. In all cases, 1 U was defined as the amount of enzymatic activity needed to digest 1 μmole of the relevant substrate described by the manufacturer in one minute.

MALDI-TOF MS. MALDI-TOF MS spectra were acquired using a Voyager DE-STR mass spectrometer (Applied Biosystems Inc.) equipped with delayed extraction technology. An acceleration voltage of 25 kV and a nitrogen laser at 337 nm was used. Mass spectra were acquired in negative linear or positive reflector mode. Either external or internal calibration was used. 20 $\mu\text{g}/\mu\text{L}$ 2,5-dihydroxybenzoic acid (DHB) in 70% ACN, 5% FA was used as matrix. Between 80 and 240 laser shots were acquired and added to generate representative spectra.

Liquid Chromatography–Tandem Mass Spectrometry (LC–MS/MS). Peptide separation was achieved by using an LC–Packings nanoflow system (LC Packings, Amsterdam, The Netherlands) equipped with a Famos autosampler. Peptides were loaded with a high flow rate of $\sim 5 \mu\text{L}/\text{min}$ onto a custom-made ~ 2 cm precolumn (75 μm i.d. fused silica with kasil-frits retaining the ODS-A C18 (YMC, Kyoto, Japan) 5/15 μm particles). Nanoflow reversed-phase HPLC was then performed

with a flow of 0.2 $\mu\text{L}/\text{min}$ through a custom-made ~ 10 cm analytical column (75 μm i.d., packed with Zorbax (Agilent Technologies, Waldbronn, Germany) 5 μm C₁₈ particles) and eluted directly into the ESI source of a Q-TOF Micro tandem mass spectrometer (Waters/Micromass, Manchester, UK) using a linear gradient from 5 to 40% acetonitrile in 0.6% acetic acid (Merck), 1% formic acid (Merck) and 0.005% heptafluorobutyric acid (ICN Biomedicals, Aurora, OH) in 35 min. Some samples were analyzed using an LC–Packings nanoflow system coupled to a Q-TOF Ultima mass spectrometer (Waters/Micromass). In this case, samples were loaded directly onto a ~ 12 cm custom-made analytical column (75 μm i.d., packed with Zorbax 5 μm C₁₈ particles) either by a Famos autosampler or by manual sample loading (“bomb loading”). Mass- and charge-dependent collision energies were used for peptide fragmentation.

Spectral Analysis and Database Searching. The theoretical masses for tryptic peptides and deglycosylated peptides of the test set of proteins were calculated using GPMW (Lighthouse data, Odense, Denmark).²⁹ The tandem mass spectra acquired during LC–MS/MS were smoothed, centroided and converted to peak-list text files using MassLynx 3.5 (Waters/Micromass). The spectral data was searched against the nonredundant databases from the National Center for Biotechnology Information (nrNCBI), SWISS-PROT and TrEMBL using the Mascot software (Matrix Sciences Ltd., Cheshire, UK). The following search parameters were used: carbamidomethylation of all cysteine residues, possible oxidation of methionine residues, possible conversion of terminal glutamine residues into pyroglutamate, at most 2 missed tryptic cleavage sites, a 0.15 Da error tolerance in both MS and MS/MS. We included two variable modifications for asparagine: N-linked GlcNAc (adding a monoisotopic mass of 203.08 Da to the mass of asparagine) and fucosylated GlcNAc (adding a monoisotopic mass of 349.14 Da). All the assigned peptides containing glycosylation sites were validated by manual interpretation of the spectra in order to ensure the presence of the NXS/T/C consensus sequon and the presence of a GlcNAc oxonium ion at m/z 204.08. A Mascot score of 25 or a sequence tag equal to or longer than four amino acids were used as acceptance criteria for peptide identification.

Results and Discussion

Overall Strategy for Determination of N-Glycosylation Sites. A global analysis of N-glycosylation sites in proteome research is challenging due to several factors. First, the signal intensities of glycopeptides are often relatively low compared to those of nonglycosylated peptides, mainly because the signal is distributed over a population of peptides species carrying different glycan structures. Second, identification of glycosylation sites using tandem mass spectrometry is complicated and fragment ion spectra of intact glycopeptides rarely allow determination of their peptide sequences. We describe a two-step analytical method designed to reduce these problems: a hydrophilic interaction chromatographic separation step designed to enrich glycosylated peptides and a partial deglycosylation step using endo- β -N-acetylglucosaminidases designed to reduce the complexity of glycopeptides and to aid in the identification of their glycosylation sites.

Capture of Glycosylated Peptides by Hydrophilic Interaction Chromatography. Hydrophilic interaction liquid chromatography (HILIC) is a chromatographic technique in which a polar stationary phase is used in combination with a less polar mobile phase.¹⁹ In this study, we have tested whether it is feasible to separate glycopeptides from more hydrophobic

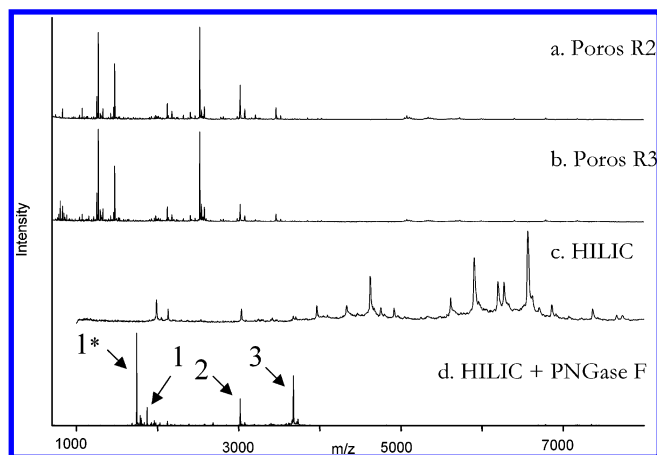


Figure 2. MALDI-TOF mass spectra of peptides from fetuin after tryptic digestion. In all cases 0.5 μ L sample solution and 0.5 μ L DHB (20 μ g/ μ L in 70%ACN/5% FA) were mixed directly on target. (a) Spectrum of peptides retained on an R2 reversed phase microcolumn. (b) Spectrum of peptides retained on an R3 reversed phase microcolumn. (c) Spectrum of peptides retained on a HILIC microcolumn. (d) Spectrum of peptides after PNGase F digestion of the sample displayed in 2c. Peaks corresponding to deglycosylated peptides are marked with numbers (the sequences of the numbered peptides are listed in Table 2). The peak marked with an asterisk corresponds to a glycopeptide containing an internal S–S bond. All spectra were acquired in positive linear mode except 2c which was acquired in negative linear mode.

nonglycosylated peptides by using a resin with a covalently bound neutral, zwitterionic sulfobetaine functional group for so-called zwitterion chromatography-hydrophilic interaction chromatography (ZIC-HILIC).³⁰ A tryptic digest of fetuin, a glycoprotein which has 3 well-characterized N-glycosylation sites,³¹ was used to evaluate HILIC. The digested protein was loaded in 80% ACN, 0.5% FA onto a ZIC-HILIC resin packed in a GELoader tip. After several washes the peptides were eluted in 0.5% FA. For a comparison, the same protein digest was loaded onto microcolumns packed with R2 and R3 reversed phase chromatographic media. Figure 2a–c show MALDI-TOF mass spectra of the peptides eluted from R2, R3, and HILIC microcolumns, respectively. The mass spectra of the peptides eluted from R2 and R3 microcolumns contain signals corresponding to several of the predicted nonglycosylated tryptic peptides from fetuin. In the case of the HILIC microcolumn,

the ion signals are mainly in a higher m/z range (3000–6000), and do not match any expected nonglycosylated tryptic peptide. Analysis of a spectrum in positive linear mode (data not shown) reveals several ion signals separated by 291 Da, which is indicative of sialylated glycopeptides. Therefore we acquired a spectrum in negative linear mode, which has been reported³² to provide strong ion signals for sialylated glycans (Figure 2c). To demonstrate that the peptides recovered from the HILIC resin were indeed glycopeptides, the sample was treated with PNGase F. As seen in Figure 2d, ion signals representing the tryptic peptides bearing the three N-glycosylation sites in fetuin are present in the MALDI-TOF mass spectrum (the sequences of these peptides are listed in Table 2). Furthermore, a fourth peptide (marked with an asterisk) was also detected which is a variant of the peptide LCPDCPLLAPLNDSR containing an internal S–S bond which seems resistant to reduction. Indeed, incomplete reduction of this bond has been reported previously.³³

Enzymatic Deglycosylation of Glycopeptides. Peptide *N*-glycosidases and endo- β -*N*-acetylglucosaminidases are the two main types of endo-acting enzymes used for deglycosylation of proteins (Figure 1). In our approach, we have used endo- β -*N*-acetylglucosaminidases to cleave the bond between the two *N*-acetylglucosamine residues in the chitobiose core (Figure 1b). This generates a peptide with either a single GlcNAc residue or a fucosylated GlcNAc residue attached to the asparagine residue. A combination of endo- β -*N*-acetylglucosaminidases was used (which will be referred to as Endo D/H) that potentially is active on most of the major N-glycan categories: 1. Endo- β -*N*-acetylglucosaminidase D (Endo D) from *S. pneumonia* is active on complex type glycans when used in combination with neuraminidase, β -galactosidase and *N*-acetyl- β -glucosaminidase.³⁴ 2. Endo- β -*N*-acetylglucosaminidase H from *S. plicatus* is active mainly on high mannose and hybrid type glycans.³⁵ Endo D was the first endo-acting glycosidase that was shown to release glycans from proteins,³⁶ but it has found only limited use as a tool for deglycosylation in recent years.

To test these enzymes for partial deglycosylation of glycopeptides, a tryptic digest of fetuin (which contain complex type glycans) was subjected to digestion by Endo D/H and the samples were analyzed by MALDI-TOF MS. As seen in Figure 3a, signals corresponding to the three tryptic glycopeptides carrying one GlcNAc residue are present in the mass spectrum, demonstrating that the deglycosylation reaction was successful.

Table 1. Summary of the Peptides Identified after LC–MS/MS Analysis of a Glycoprotein Mixture, Either after Endo D/H Treatment (Endo D/H) or after HILIC Microcolumn Separation Followed by Endo D/H Treatment (HILIC-Endo D/H)^a

protein name	accession no.	known glycosylation sites	endo D/H		HILIC-endo D/H	
			nonglycosylated peptides/glycopeptides	glycosylation sites	nonglycosylated peptides/glycopeptides	glycosylation sites
fetuin	P12763	3	7/3	3	1/4	3
α -1-acid glycoprotein 1	P02763	5	7/3	2	0/3	2
α -1-acid glycoprotein 2	P19652	5	1/0	0	0/1	1
ovalbumin	P01012	1	11/0	0	1/0	0
RNase B	P00656	1	3/0	0	0/0	0
ovomucoid	P01005	5	3/1	1	0/0	0
total		20	32/7	6	2/8	6

^a 10 pmol of each sample was separated on a nano-LC system coupled to a Q-TOF Micro mass spectrometer. Data from MS/MS spectra was searched against NCBI nr using Mascot (Matrix Sciences Ltd.) with GlcNAc (203.079) and fucosylated GlcNAc (349.137) as variable modifications and tryptic constraints. The identified proteins and the number of known N-glycosylation sites in these are listed. The number of identified nonglycosylated peptides/glycopeptides and the number of glycosylation sites covered are also displayed. The number of identified glycopeptides and glycosylation sites are different in some cases due to the presence of more than one peptide containing the same site caused by incomplete tryptic cleavage. All identified peptides gave a score above 25.

Table 2. Glycopeptides Identified in the Test Glycoprotein Mixture^a

protein name	peptide no.	sequence
fetuin (P12763)	1	LCPDCPLLAPL n DSR KLCPCDLLAPL n DSR NAES n GSYLQLVEISR [§]
	2	VVHAVEVALATF n AESN#GSYLQLVEISR
	3	RPTGEVYDIEIDTLETTCHVLDP TPLAn CSVR PVPIT n ATLDQITGK [§]
α-1-acid glycoprotein 1 (P02763)	4	LVPVPIT n ATLDQITGK [§] QDQCIY n TTYLVNQ R
	5	SVQEIQATFFYFT Pn K SVQEIQATFFYFT Pn KTEDTIFLR
		PVPIT n ATLDR [§]
α-1-acid glycoprotein 2 (P19652)		LVPVPIT n ATLDR [§] qn QCFY N SSYLNVQ R qn QCFY N SSYLNVQ R
		SIEFGT n ISK [§]
		CNFC n AVVES N GTLTLSHF GK
ovomucoid precursor (P01005)		

^a 10 different glycosylated peptides were identified when searching with tryptic constraints. An additional six peptides, marked with §, were identified when searching with semi-tryptic constraints. The lowercase bold letters indicate the Mascot annotation of posttranslational modifications: **q** denotes pyroglutamic acid, **m** denotes oxidized methionine residues, **n** denotes GlcNAc-modified asparagine residues. The asparagine residues followed by # are the glycosylation sites found within the glycosylation sequon. All identified tryptic peptides gave a Mascot score above 25. In the case of semi-tryptic peptides, we manually interpreted the spectra and only accepted assignments based on sequence tags equal to or longer than four amino acids. The numbering of the peptides refers to those used in Figures 2–4.

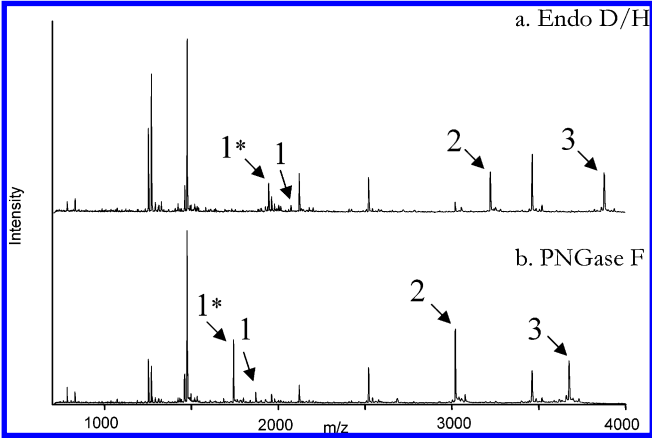


Figure 3. MALDI-TOF mass spectra of tryptic peptides from fetuin after digestion by endoglycosidases. (a) Spectrum after digestion with Endo D/H. (b) Spectrum after digestion with PNGase F. In both cases 0.5 μ L sample (0.5 pmol) and 0.5 μ L DHB (20 μ g/ μ L in 70% ACN/5% FA) were mixed directly on target. Peaks corresponding to deglycosylated peptides (with a mass shift of +0.98 Da in the case of PNGase F and +203.08 Da in the case of Endo D/H) are marked with numbers (the sequences of the numbered peptides are listed in Table 2). The peaks marked with asterisks correspond to a glycopeptide containing an internal S–S bond.

To compare the relative intensities of the GlcNAc-modified peptides with the corresponding deamidated peptides formed after hydrolysis with PNGase F, a tryptic digest of fetuin was digested with PNGase F and the resulting peptides were analyzed by MALDI-TOF MS. The resulting spectrum (Figure 3b) shows that the intensities of the signals from the peptides bearing the glycosylation sites are in the same range as in Figure 3a, when compared to the base peaks in the spectra. These results suggest that the single GlcNAc residues attached to these peptides do not reduce the signal intensity in MALDI-TOF MS analysis. However, it must also be taken into account

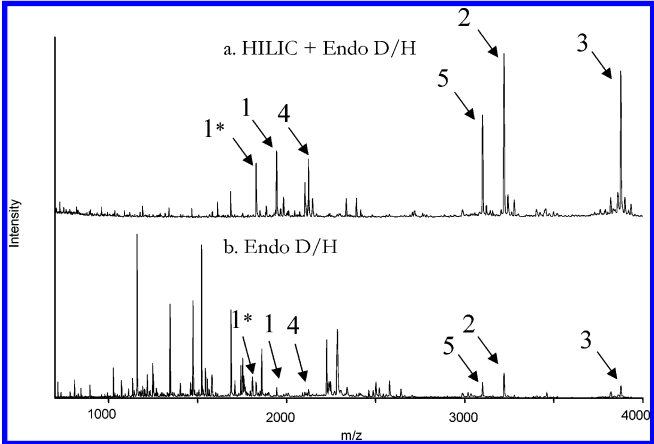


Figure 4. MALDI-TOF mass spectra of tryptic peptides from a glycoprotein mixture (a) Spectrum of HILIC eluent digested with Endo D/H. (b) Spectrum of an Endo D/H digest. In both cases 0.5 μ L tryptic digest (1 pmol) and 0.5 μ L DHB (20 μ g/ μ L in 70% ACN/5% FA) was mixed directly on target. Peaks corresponding to deglycosylated peptides (with GlcNAc attached) are marked with numbers (the sequences of the numbered peptides are listed in Table 2). The peaks marked with asterisks correspond to a glycopeptide containing an internal S–S bond.

that PNGase F converts asparagine into aspartate, which may reduce the signal intensity in positive-mode MALDI-TOF MS. To ensure that Endo D/H was active also on high-mannose glycans, we performed a digestion of intact RNase B, a low molecular weight protein that contains one high-mannose type glycosylation site. Analysis by MALDI-TOF MS revealed a reduced molecular weight after digestion, which corresponds to the expected molecular weight of the RNase B protein with one GlcNAc residue attached (data not shown). To assess the feasibility of applying our approach to a more complex sample we used a mixture of four glycoproteins, i.e., fetuin, α-1-acid

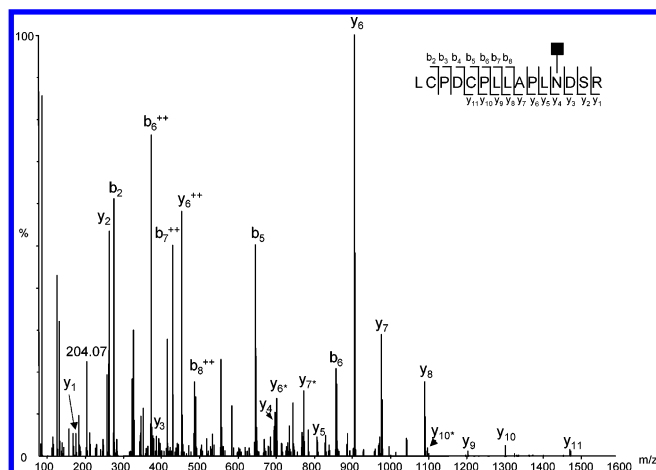
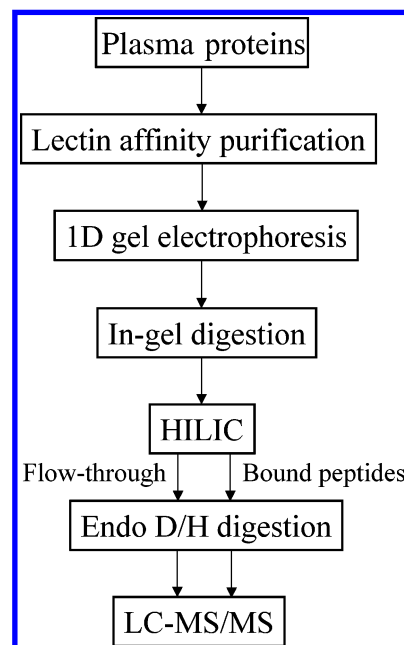


Figure 5. LC-ESI tandem mass spectrum of the peptide with a $[M+2H]^{2+}$ precursor ion at m/z 972.46 acquired on a Q-TOF Micro. The CID fragment ion spectrum has been assigned to the peptide LCPDCPLLAPLNDNR from fetuin with a GlcNAc residue attached to the asparagine. The y and b ion series are indicated and the y-ions with an asterisk indicate the neutral loss of GlcNAc. From the fragmentation pattern of the glycopeptide heteroconjugate it is possible to unambiguously assign the glycosylation site to the asparagine residue. The oxonium ion of GlcNAc at m/z 204.08 serves as a diagnostic ion for glycopeptides.

glycoprotein, RNase B and ovalbumin. In an initial experiment, 10 pmole of a tryptic digest of the glycoprotein mixture was applied to a HILIC microcolumn and subsequently digested by Endo D/H. As a comparison, the same sample was also digested with Endo D/H without prior HILIC separation and both samples were analyzed by MALDI-TOF MS (Figure 4). As seen, the complexity of the peptide mass spectrum is significantly reduced when HILIC separation is included (Figure 4a) in comparison with Endo D/H digestion alone (Figure 4b). As indicated in Figure 4a, most of the high intensity signals in this spectrum correspond to glycopeptide ions. The corresponding glycopeptide signals in Figure 4b are generally of much lower relative intensity in comparison with the nonglycosylated peptides, indicating that the HILIC microcolumn separation provided an enrichment of glycopeptides from the peptide/glycopeptide mixture.

PNGase F is capable of releasing almost all types of asparagine-linked glycans and in the process asparagine is converted to aspartate with a concomitant mass increase of 0.98 Da (Figure 1a). While this mass increase is more than sufficient for most mass spectrometers with good mass accuracy (e.g., most modern Q-TOF instruments) to distinguish between asparagine and aspartate it may be insufficient in the case of mass spectrometers with lower mass accuracy (e.g., some ion trap instruments). Furthermore, it can become difficult to interpret the spectra in the case of partial site occupancy because the isotope pattern of the nonglycosylated peptide overlaps that of the glycosylated peptide. In a more sophisticated approach, PNGase F is used in the presence of ^{18}O labeled water, leading to the incorporation of ^{18}O at the asparagine residue. This method has been widely used for proteins separated by gel-electrophoresis,^{23,24} and recently in a large-scale study of lectin affinity purified glycopeptides.¹² The drawbacks of the PNGase F ^{18}O -labeling is that it cannot eliminate the possibility of spontaneous deamidation of asparagines residues and in some cases two ^{18}O are incorporated.²⁵ Even with incorporation of ^{18}O the overlap of isotope

Scheme 1 Overview of the Approach Used for Analysis of N-Glycosylation Sites in Proteins from Human Plasma



patterns is problematic with low occupancy glycosylation sites where the signal from the deglycosylated peptide can be overshadowed by the nonglycosylated peptide. Obviously, the use of ^{18}O labeling reduces the likelihood of false positive identification of glycosylation sites since the spontaneous deamidation should take place after the PNGase F treatment, but it does not match the undisputable direct evidence for N-glycosylation achieved by the Endo D/H treatment described here.

Tandem Mass Spectrometry of Glycopeptide Heteroconjugates. To test the applicability of our method for analysis of glycosylation sites using LC-MS/MS followed by database searching, we used the same glycoprotein mixture as described above. Briefly, the Endo D/H digested samples (with or without prior HILIC microcolumn separation) were separated on a nano-LC system and analyzed on-line with an ESI Q-TOF tandem mass spectrometer. A total of 7 tryptic glycopeptides and 32 nonglycosylated peptides were found in the Endo D/H digested sample when the MS/MS data was searched against the nrNCBI database using Mascot with N-linked GlcNAc (adding a monoisotopic mass of 203.08 Da) and fucosylated GlcNAc (adding a monoisotopic mass of 349.14 Da) as variable modifications (Table 1). In the sample that was pre-fractionated by HILIC prior to the Endo D/H digestion a total of 10 tryptic peptides were identified, of which 8 were glycopeptides. Thus, a similar number of glycopeptides were identified, but the glycopeptide fraction of the identified peptides was considerably larger when HILIC was used. This is in agreement with the results obtained from MALDI-TOF MS analysis of the same samples (Figure 4). The identified glycopeptides were mainly from fetuin and two variants of α -1-acid glycoprotein, but the single glycopeptides in ovalbumin and RNase B were not identified. It may be that these peptides do not bind to the HILIC resin, but it is also possible that they are lost in some other step of the sample preparation. The tryptic glycopeptide from RNase B is very small (677 Da, including a GlcNAc residue) and may not be detected after LC-MS/MS. Apart from the proteins included in the glycoprotein mixture, some peptides (including one glycopeptide) matched ovomucoid that is

Table 3. Glycoproteins Identified in Plasma with Glycopeptides Detected in HILIC-bound Fractions^a

entry name	accession no.	entry description	glycopeptide
A1AG_HUMAN	P02763	α-1-acid glycoprotein 1 precursor (5,5)	QDQCIYn#TTYLVNVQR qIPLCAnLVVPVITN#ATLDQITGK [§]
A1AH_HUMAN	P19652	α-1-acid glycoprotein 2 precursor (5,5) ⇨	qnQCFYN#SSYLVNVQR SVQEIQATFFYFTPn#K SVQEIQATFFYFTPn#KTEDTIFLR qIPLCAnLVVPVITN#ATLDR [§] FFYFTPn#KTEDTIFLR [§] LVVPVITn#ATLDR [§] YLGn#ATAIFFLPDEGK KLPPGLLAn#FTLLR LPPGLLAn#FTLLR VSn#QTLSTFTVLQDVPVR (+fuc) SLGnVn#FTVSAEALESQELCGTEVPSVPEHGR SLGnVn#FTVSAEALESQELCGTEVPSVPEHGRK LSLHRPALEDLLGSEAn#LTCTLTGLR* PALEDLLGSEAn#LTCTLTGLR [§] LGACn#DTLQQLMEVFK Fn#SSYLQGTNQITGR (pot.) YDFn#SSmLYSTAK (pot.) FEVDSPVYn#ATWSASLK (pot.) VnQNLVYESGSLN#FSK (pot.) VFHIn#ESWVLLTPK (pot.) IDSTGn#VTNELR (pot.) LGHCNPDVnVnGEFSSGPN#VSDK (pot.) AFGSnPN#LTK (pot.) ELHHLQEQn#VSNFLDK En#LTAPGSDSAVFEEQGTTR mDGSAn#VTCINSR (pot.) SPDVIn#GSPISQK (pot.) ISEEn#ETTCYmGK (pot.) nGN#WTEPPQCK [§] (pot.) IPCSQPPQIEHGtIn#SSR (pot.) – (fuc-only) LSDLSIn#STECCLHVHCR (pot.) LAn#LTQGEDQYYLR (+ fuc) qVPAHFVALn#GSK TVLTPATnHMGn#VTFIPANR GLn#VTLSTGR (pot.) (+fuc) n#LTVSVHVSPVEGLCLAGGGGLAQQLVLPAGSAR [§] (pot.) InNDFNYEFYN#STWSYVK (pot.) LQnNENN#ISCVLR (+fuc) LDAPTnLQFVN#ETDSTVLVR (pot.) DQCIVDDITYnVn#DTFHK (pot.) EEQFn#STFR (n.a.) (+fuc) SWPAVGn#CSSALR MVSHHn#LTTGATLINEQWLLTTAK VVLHPn#YSQVDIGLIK nLFLn#HSEN#ATAK (+ fuc) GLCVn#ASAVSR – (fuc-only) VLSn#NSDANLELINTWVAK (+fuc) VGQLQLShn#LSLVLPQNLK (+fuc) En#ISDPTSPLR IIVPLnNREN#ISDPTSPLR (+fuc) LPTQn#ITFQTESSVAEQAEFQSPK (pot.) IYSGILn#LSDITK IYPGVDFGGEELn#VTFVK LnAENN#ATFYFK (pot.) VTQVYAEn#GTVLQGSTVASVYK (pot.) HAn#WTLTPLK (pot.) AAIPsALDTn#SSK (n.a.) ALGFEn#ATQALGR (n.a.) TVIRPFYLTn#SSGVD (n.a.) ESVTDHVnLITPLEKPLQn#FTLCFR ESVTDHVnLITPLEKPLQn#FTLC [§] LITPLEKPLQn#FTLCFR [§] EKPLQn#FTLCFR [§] PLQn#FTLCFR [§] SRYPHKPEIn#STTHPGADLQENFCR WVLTAAH CLLYPPWDKn#FTENDLLVR YPHKPEIn#STTHPGADLQENFCR CLLYPPWDKn#FTENDLLVR [§] VVn#STTGPGELR (pot.) n#GSLFAFR n#ISDGFDFGIPDNVDAALPAHSYSGR DIVEYn#DSNGSHVLQGR
A2MG_HUMAN	P01023	α-2-macroglobulin precursor (8,8)	
ALC1_HUMAN	P01876	Ig α-1 chain C region (2,2)	
ANT3_HUMAN	P01008	antithrombin-III precursor (4,4)	
APB_HUMAN	P04114	apolipoprotein B-100 precursor (19,19) ⇨	
ATRN_HUMAN	O75882	atractin precursor (26,26) ⇨	
C4BB_HUMAN	P20851	C4b-binding protein β chain precursor (5,5) ⇨	
CBP8_HUMAN	P22792	carboxypeptidase N 83 kDa chain (8,8) ⇨	
CERU_HUMAN	P00450	ceruloplasmin precursor (7,4)	
CFAH_HUMAN	P08603	complement factor H precursor (9,8)	
CFAL_HUMAN	P05156	complement factor I precursor (6,6)	
CLUS_HUMAN	P10909	clusterin precursor (6,6)	
CO2_HUMAN	P06681	complement C2 precursor (8,8)	
CO3_HUMAN	P01024	complement C3 precursor (3,3)	
CO4_HUMAN	P01028	complement C4 precursor (4,4)	
CO7_HUMAN	P10643	complement component C7 precursor (2,2)	
FHR1_HUMAN	Q03591	complement factor H-related protein 1 p (2,2)	
FINC_HUMAN	P02751	fibronectin precursor (10,6) ⇨	
GC2_HUMAN	P01859	Ig γ-2 chain C region (1,0)	
HEMO_HUMAN	P02790	hemopexin precursor (5,5)	
HPT1_HUMAN	P00737	haptoglobin-1 precursor (4,4) ⇨	
IBP3_HUMAN	P17936	insulin-like growth factor binding protein (3,3) ⇨	
IC1_HUMAN	P05155	plasma protease C1 inhibitor precursor (6,6)	
IGJ_HUMAN	P01591	immunoglobulin J chain (1,1)	
ITH4_HUMAN	Q14624	inter-α-trypsin inhibitor heavy chain H4 (4,4)	
KAL_HUMAN	P03952	plasma kallikrein precursor (5,5) ⇨	
KNQ_HUMAN	P01042	kininogen precursor (4,4)	
PON1_HUMAN	P27169	serum paraoxonase/arylesterase 1 (4,3)	
Q08380	Q08380	MAC-2 binding protein precursor (7,0)	
SAMP_HUMAN	P02743	serum amyloid P-component precursor (2,1)	
THRB_HUMAN	P00734	prothrombin precursor (4,3)	
TSP1_HUMAN	P07996	thrombospondin 1 precursor (4,4)	
VTNC_HUMAN	P04004	vitronectin precursor (3,3)	
ZA2G_HUMAN	P25311	zinc-α-2-glycoprotein precursor (4,3)	

^a 62 different glycosylated peptides from 37 different proteins were identified when searching with tryptic constraints. An additional 12 peptides, marked with §, were identified when searching with semi-tryptic constraints. Only one peptide, marked with an asterisk, was also found among the HILIC flow-through fractions. For each SWISS-PROT entry, the accession number and a descriptive name of the protein is given (listed in parentheses are the number of putative N-glycosylation sequons and SWISS-PROT annotated glycosylation sites, respectively). Proteins for which all identified peptides were from the HILIC-bound fractions are marked with ⇨. The lowercase bold letters indicate the Mascot annotation of posttranslational modifications: **q** denotes pyroglutamic acid, **m** denotes oxidized methionine residues, **n** denotes the modified asparagine residue. The asparagine residues followed by # are the glycosylation sites found within glycosylation sequons. On the basis of the information available in SWISS-PROT, some glycopeptides are marked as potential (pot.) or not annotated (n.a.). In some cases both a GlcNAc- and fucosylated GlcNAc-modified peptide was identified (+fuc) and in two cases only the fucosylated GlcNAc-modified peptides (fuc-only). In all cases, we manually interpreted the spectra and only accepted assignments based on sequence tags equal to or longer than 4 amino acids.

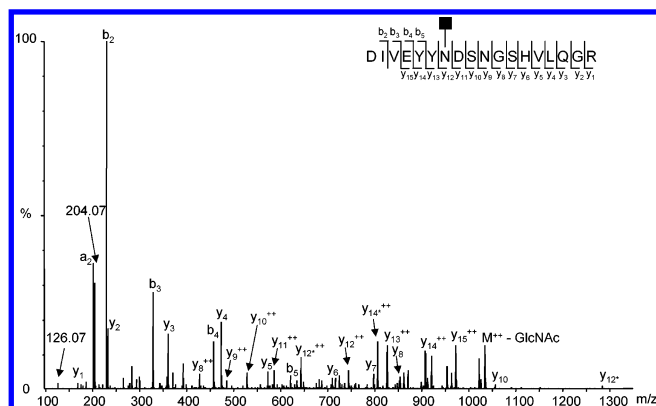


Figure 6. LC-ESI tandem mass spectrum of a peptide with a $[M+3H]^{3+}$ precursor ion at m/z 757.05 acquired on a Q-TOF Ultima. The fragment ion spectrum has been assigned to the peptide DIVEYYNDSNGSHVLQGR (with a GlcNAc residue attached) from zinc- α -2-glycoprotein (ZA2G_HUMAN) that has two putative glycosylation sites. The y and b ion series are indicated and the y-ions with an asterisk indicate the neutral loss of GlcNAc. Analysis of the fragment ion series shows that NDS is the main glycan attachment site. Note the signal from the GlcNAc oxonium ion at m/z 204.08 and the GlcNAc fragment ion at m/z 126.08.

known to be co-purified with ovalbumin in commercial preparations.³⁷

For analysis of complex mixtures, it is important to identify the amino acid sequences of the peptides through CID tandem mass spectra followed by database searching but this poses a problem for most glycopeptide heteroconjugates. As noted by Jebanathirajah,¹⁵ N-glycans are labile and a very low amount of collision energy is needed to cleave the glycosidic bonds. Tandem mass spectra of glycopeptides are obviously very complex since they contain fragments of both the glycan and the peptide moieties. Since peptide fragmentation requires more energy than the fragmentation of glycans, peptide fragment ions often appear as low intensity signals generated by secondary fragmentation processes after the primary fragmentation of carbohydrate moieties. However, by reducing the size of the glycan down to only one or two carbohydrate residues it is possible to simplify the fragmentation patterns. The trial runs indicated that the mass- and charge-dependent collision energies used for unmodified peptides were suitable for the peptides carrying a single N-linked GlcNAc residue. Thus, the N-linked GlcNAc seems to be more stable than O-linked GlcNAc residues for which a lower collision energy is used to obtain interpretable MS/MS spectra.³⁸

An example of a fragment ion spectrum from a GlcNAc-modified peptide is shown in Figure 5. As displayed in this spectrum, several peptide fragment ions contain the GlcNAc residue. Furthermore, the GlcNAc oxonium ion at m/z 204.08 is present. A survey of the fragment ion spectra obtained from the glycoprotein mixture analysis revealed that this oxonium ion is prominent in most of the spectra containing the GlcNAc modification. As previously described, the 204.08 m/z -value is a good indicator for the GlcNAc oxonium ion, but it should be noted that a number of potentially interfering peptide fragment ions may be present at a similar m/z value.¹⁵ We screened the LC-MS/MS data to find unassigned peptides with a prominent m/z 204.08 signal in their fragment ion spectra, and as indicated in Table 2 we uncovered six unassigned peptides containing glycosylation sites. Four of these peptides stem from

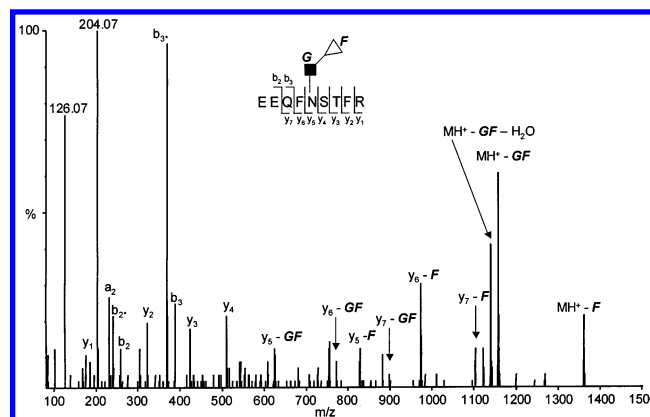


Figure 7. LC-ESI tandem mass spectrum of a peptide with a $[M+2H]^{2+}$ precursor ion at m/z 753.88 acquired on a Q-TOF Micro. The fragment ion spectrum has been assigned to the peptide EEQFNSTFR from Ig γ -2 chain C region (GC2_HUMAN) carrying a core fucosylated GlcNAc at the glycosylation site. The y and b ion series are indicated and the y-ions marked with F and GF indicate the neutral loss of fucose and fucosylated GlcNAc, respectively. The fragmentation pattern for the core-fucosylated peptide is more complex than that of the peptide with only a GlcNAc residue attached. The signal from the GlcNAc oxonium ion at m/z 204.08 is prominent in the spectrum and a GlcNAc fragment ion at m/z 126.08 is also present.

the N-termini of α -1-acid glycoproteins 1 and 2. These peptides do not correlate with the predicted signal peptide cleavage site, but could possibly be the result of exopeptidase activity (e.g., aminopeptidase). These N-terminal glycopeptides were also found in a recent independent study.³⁹ It is difficult to establish a cause for the presence of these semi-tryptic peptides but further investigations (using ¹⁸O-labeling during digestion) could possibly help to distinguish between the various possibilities, such as in-vivo or ex-vivo enzymatic protein processing, in-source fragmentation or tryptic miscleavage. In addition to the 204.08 oxonium ion, several GlcNAc fragment ions were observed at m/z 126.1, 138.1, 168.1 and 186.1. While not as intense as the oxonium ion itself, these fragments may potentially be used in combination with the m/z 204.08 ion as markers of glycopeptides to increase the confidence of glycopeptide identifications.

Analysis of N-Glycosylation Sites in Human Plasma. Since our approach to identify glycosylation sites in the glycoprotein mixture was partially successful, we proceeded to use our strategy for the analysis of N-glycosylation sites in human plasma. Several recent reports have been published in which mass spectrometry has been utilized in conjunction with either 2D-electrophoresis or chromatographic separation for large-scale identification of proteins in plasma.^{40–42} In the context of proteomics, analysis of human plasma is a very challenging task since it is dominated by high abundance proteins (e.g., serum albumin and immunoglobulins), and several methods for the removal of these have been described.^{42,43} Moreover, the dynamic range of protein concentration in plasma spans up to 10 orders of magnitude.⁴⁴ Due to these problems, we decided to include lectin chromatography as an initial glycoprotein enrichment step (Scheme 1). Plasma was incubated with Con A lectin and the bound proteins were analyzed by SDS-PAGE, followed by tryptic in-gel digestion. The digests were applied to HILIC microcolumns and both the flow-through and the bound peptides were digested with Endo D/H. These samples were subsequently analyzed by LC-MS/MS and

Table 4. Proteins Identified in Plasma Based on Non-glycosylated Peptides^a

entry name	entry description
A1BG_HUMAN	α -1B-glycoprotein precursor (4, 4)
ALBU_HUMAN	serum albumin precursor (0, 0) *
ALC2_HUMAN	Ig α -2 chain C region (4, 4)
ALS_PAPHA	insulin-like growth factor binding protein complex acid labile chain precursor (6, 6)
AMBP_HUMAN	AMBP protein precursor (3, 3)
APA1_HUMAN	apolipoprotein A-I precursor (0, 0)
APE_HUMAN	apolipoprotein E precursor (0, 0)
APOD_HUMAN	apolipoprotein D precursor (2, 2)
APOH_HUMAN	β -2-glycoprotein I precursor (4, 4)
APOM_HUMAN	apolipoprotein M (2, 1)
C1QA_HUMAN	complement C1q subcomponent, A chain precursor. (1, 1)
C1QB_HUMAN	complement C1q subcomponent, B chain precursor (0, 0)
C1QC_HUMAN	complement C1q subcomponent, C chain precursor (0, 0) *
C1R_HUMAN	complement C1r component precursor (4, 4)
C1S_HUMAN	complement C1s component precursor (3, 2) *
C4BP_HUMAN	C4b-binding protein α chain precursor (3, 3)
CD5L_HUMAN	CD5 antigen-like precursor (0, 0)
CFAB_HUMAN	complement factor B precursor (5, 4)
CO5_HUMAN	complement C5 precursor (4, 4)
CO8B_HUMAN	complement component C8 β chain precursor. (4, 2)
CO8G_HUMAN	complement component C8 γ chain precursor (0, 0)
CONA_CANVI	concanavalin A (2, 0) *
EBAG_STRPL	endo-beta-N-acetylglucosaminidase H precursor (3, 0) #
ECM1_HUMAN	extracellular matrix protein 1 precursor (3, 3)
FA5_HUMAN	coagulation factor V precursor (37, 26)
FBL1_HUMAN	fibulin-1 precursor. (3, 3)
FCN3_HUMAN	ficolin 3 precursor (1, 1)
FHR2_HUMAN	complement factor H-related protein 2 precursor (1, 1)
FHR5_HUMAN	complement factor H-related protein 5 precursor (2, 2)
GC1_HUMAN	Ig γ -1 chain C region (1, 1) *
GC4_HUMAN	Ig γ -4 chain C region (1, 0) *
HBB_HUMAN	hemoglobin β chain (0, 0)
HPT2_HUMAN	haptoglobin-2 precursor (4, 4)
HPTR_HUMAN	haptoglobin-related protein precursor (3, 0)
HRG_HUMAN	histidine-rich glycoprotein precursor (4, 4) *
K1CI_HUMAN	keratin, type I cytoskeletal 9 (3, 0)
K22E_HUMAN	keratin, type II cytoskeletal 2 epidermal (2, 0)
K2C1_HUMAN	keratin, type II cytoskeletal 1 (3, 0)
KAC_HUMAN	Ig κ chain C region (0, 0) *
KV2A_HUMAN	Ig κ chain V-II region Cum. (0, 0)
KV4A_HUMAN	Ig κ chain V-IV region Len. (0, 0)
LAC_HUMAN	Ig λ chain C regions (0, 0) *
LBP_HUMAN	lipopolysaccharide-binding protein precursor (5, 4)
LV1C_HUMAN	Ig λ chain V-I region NEW. (0, 0)
LV1E_HUMAN	Ig λ chain V-I region NEWM. (0, 0)
LV1F_HUMAN	Ig λ chain V-I region WAH. (0, 0)
LV4C_HUMAN	Ig λ chain V-IV region Hil. (0, 0)
MAS2_HUMAN	Mannan-binding lectin serine protease 2 precursor (1, 0)
MUC_HUMAN	Ig μ chain C region (5, 5) *
PROP_HUMAN	properdin precursor (1, 1) #
PRTS_HUMAN	vitamin K-dependent protein S precursor. (3, 3)
TRYP_PIG	trypsin precursor (0, 0)
TTHY_HUMAN	transthyretin precursor (1, 0) *

^a This table is based on LC-MS/MS analysis of bound and flow-through fractions from HILIC of in-gel digested samples. Protein identified based on peptides from both flow-through and HILIC-bound fractions are marked with asterisks and proteins with peptides found only in the HILIC-bound fractions are marked with #. For each SWISS-PROT entry, a descriptive name of the protein is given and, in parentheses, the number of putative N-glycosylation sequons and SWISS-PROT annotated glycosylation sites, respectively.

from database searching, using tryptic constrains, 62 tryptic glycopeptides from 37 glycoproteins were identified (Table 3). Strikingly, all the glycopeptides were found in the HILIC-bound fractions, again showing the effectiveness of HILIC for glycopeptide enrichment. Only one peptide (LSLHRPALEDLLLG-SEANLTCTLTGLR) bearing a GlcNAc residue was identified among the flow-through fractions and this peptide from Ig α -1 chain C region (P01876) was also found in the corresponding HILIC-bound fraction. For 10 of the proteins all identified peptides were from the HILIC-bound fractions, indicating that these may be of low abundance (Table 3). A large number of nonglycosylated peptides were identified in the flow-through

fractions (data not shown). Thus, it was possible to increase the confidence in the protein identifications by the dual analysis of the flow-through and bound fractions.

Even though several of the identified glycoproteins are highly abundant (e.g., immunoglobulins, transferrin and haptoglobin), we also identified glycopeptides from some less abundant glycoproteins (e.g., some complement proteins).⁴⁴ Four of the identified glycopeptides stem from two proteins (MAC-2 binding protein and Ig gamma-2 chain C region) that were not annotated as glycoproteins in SWISS-PROT even though they contain potential glycosylation sequons (Table 3). In all cases, a glycosylation NXS/T/C sequon was present in the identified

peptides. Currently, it is not possible to constrain the glycosylation modification so that the Mascot search engine only considers a peptide when an asparagine appears in the consensus sequon. As listed in Table 3, Mascot returned an asparagine outside of the glycosylation sequon as the glycosylated residue in 15 cases for the tryptic peptides. Through manual inspection it was possible to identify the asparagine residue in the consensus sequon as the correct glycosylation site in 10 out of the 15 cases. In the remaining cases, it was not possible to distinguish between different asparagines, but experience suggests that the glycosylation site most likely appears within the glycosylation sequon. In a few cases, more than one consensus sequon was found within a single GlcNAc-peptide and Figure 6 shows the fragment ion spectrum of one such example, the peptide DIVEYYNDSNGSHVLQGR from zinc- α -2-glycoprotein. Analysis of the y-ion series in this spectrum enabled us to locate the main attachment sequon in this peptide to NDS. However, since glycosylation site occupancy may show considerable heterogeneity, it is possible that a small population of the peptides is modified at the other consensus sequon.

There are only a few examples of peptides with missed trypsin cleavage sites and usually these can be attributed to multiple consecutive basic residues or glycosylation sites proximal to the expected cleavage site. Another 12 glycopeptides were found when searching with semi-tryptic constraints and these are also listed in Table 3. Two of these peptides define the N-termini of α -1-glycoprotein 1 and 2 after cleavage of the signal peptides. The two peptides from complement C4 (P01028) and Ig α -1 chain C region (P01876) were not found in the initial searches because trypsin was not expected to cleave at an arginine followed by a proline. In a number of cases, shorter fragments of fully tryptic peptides were found. In particular, four different fragments of the tryptic peptide from serum amyloid P-component were detected. We compared the retention times of the four peptides with that of the full-length tryptic peptide, and we found that their elution times did not overlap. Thus, it was concluded that the appearance of the different ions was not due to in-source fragmentation.

Information concerning fucosylation of the core GlcNAc is interesting from many aspects. For example, it may be valuable in clinical applications since core-fucosylation has been correlated with cancer development.⁴⁵ It has previously been shown that both Endo D and Endo H display enzymatic activity on core-fucosylated glycopeptide substrates,^{34,46} and we identified 15 peptides carrying a fucosylated GlcNAc moiety (Table 3). In most of these cases, the corresponding nonfucosylated, GlcNAc-modified glycopeptide was also detected, but two peptides (from complement factor H and insulin like growth factor binding protein) were only identified as core-fucosylated peptides. The fucosylated peptides show fragmentation patterns that are different from those of GlcNAc-modified peptides as illustrated in Figure 7. The fucose is more labile than the N-linked GlcNAc, and it is not possible to assign the attachment site from the fragmentation spectrum, but we assume that the fucose residue stems from the N-glycan. However, it should be noted that O-fucosylation of serine and threonine occur in some rare cases.⁴⁷

Even though fucosylation of the proximal GlcNAc in the core is a common theme among glycoproteins we only found a limited number of glycosylation sites bearing this modification. One possible explanation for this observation could be that we have used a lectin (Con A) that preferentially binds to high-

mannose type glycans and these are normally not fucosylated. Another possible explanation could be that the labile fucose residues are lost due to skimmer fragmentation prior to precursor ion selection.

Another 53 proteins were identified based on nonglycosylated peptides present in the flow-through or bound fractions from the HILIC microcolumns (Table 4). Among these, 40 proteins were identified based on peptides in the flow-through fractions, 11 proteins were identified from both the flow-through and the HILIC-bound fractions, and two were only found in the HILIC-bound fractions. 36 of the proteins were annotated as glycoproteins or contained potential glycosylation sequons (Table 4). The remaining 17 proteins did not contain N-glycosylation sequons, but most of these were either high abundance proteins such as serum albumin or proteins used in the sample preparation (i.e., trypsin and glycosidases). The lack of glycosylation sequons does not necessarily indicate unspecific binding to the Con A column since, in a few cases, the retrieved entries are subcomponents of a larger protein. For example, the protein complement 1q has a single glycosylation site but has three different subcomponent entries in SWISS-PROT (two of which do not contain a glycosylation sequon). It is also possible that some of the proteins show affinity for the carbohydrate-binding lectin as a result of O-glycosylation or nonenzymatic glycation.

Conclusion

The results presented show that HILIC separation in combination with endoglycosidase digestion can be used for the analysis of N-glycosylation sites in complex protein mixtures. Altogether, 62 glycosylation sites were identified among 37 glycoproteins from human plasma. This is, to our knowledge, the first report of the use of endo- β -N-acetylglucosaminidases for comprehensive identification of N-glycosylation sites in proteomics. Con A lectin chromatography and HILIC was used for enrichment of glycoproteins and glycopeptides, respectively. However, it is apparent that the specificity of these chromatographic techniques has certain limitations. By using the two separation steps in series, it was possible to partially overcome the lack of specificity. The results show that it is possible to increase the success rate for identification of glycoproteins in proteomics, but also that some glycosylated peptides and low abundance glycoproteins are missed. Further investigations will be carried out to develop this methodology for improved detection and characterization of low abundance proteins in proteomics research projects.

Acknowledgment. P.H. is supported by a long-term fellowship from the Federation of European Biochemical Societies. F.E. was supported by a postdoctoral fellowship from the Basque government. We thank www.ionsource.com for hosting an informative website and providing templates for the making of Figure 1. Financial support for this study was provided by the Danish Research Agency through the Danish Biotechnology Instrument Centre.

References

- (1) Lis, H.; Sharon, N. *Eur. J. Biochem.* **1993**, *218*, 1–27.
- (2) Daniels, M. A.; Hogquist, K. A.; Jameson, S. C. *Nat. Immunol.* **2002**, *3*, 903–910.
- (3) Dwek, M. V.; Ross, H. A.; Leatham, A. J. *Proteomics* **2001**, *1*, 756–762.
- (4) Freeze, H. H. *Glycobiology* **2001**, *11*, 129R–143R.
- (5) Bause, E.; Hettkamp, H. *FEBS Lett.* **1979**, *108*, 341–344.

- (6) Bause, E.; Legler, G. *Biochem. J.* **1981**, 195, 639–644.
- (7) Apweiler, R.; Hermjakob, H.; Sharon, N. *Biochim. Biophys. Acta* **1999**, 1473, 4–8.
- (8) Kornfeld, R.; Kornfeld, S. *Annu. Rev. Biochem.* **1985**, 54, 631–664.
- (9) Mortz, E.; Sareneva, T.; Julkunen, I.; Roepstorff, P. *J. Mass. Spectrom.* **1996**, 31, 1109–1118.
- (10) Stahl, B.; Klabunde, T.; Witzel, H.; Krebs, B.; Steup, M.; Karas, M.; Hillenkamp, F. *Eur. J. Biochem.* **1994**, 220, 321–330.
- (11) Sutton, C. W.; O'Neill, J. A.; Cottrell, J. S. *Anal. Biochem.* **1994**, 218, 34–46.
- (12) Kaji, H.; Saito, H.; Yamauchi, Y.; Shinkawa, T.; Taoka, M.; Hirabayashi, J.; Kasai, K.; Takahashi, N.; Isobe, T. *Nat. Biotechnol.* **2003**, 21, 667–672.
- (13) Zhang, H.; Li, X. J.; Martin, D. B.; Aebersold, R. *Nat. Biotechnol.* **2003**, 21, 660–666.
- (14) Carr, S. A.; Huddleston, M. J.; Bean, M. F. *Protein Sci.* **1993**, 2, 183–196.
- (15) Jebanathirajah, J.; Steen, H.; Roepstorff, P. *J. Am. Soc. Mass. Spectrom.* **2003**, 14, 777–784.
- (16) Schindler, P. A.; Settineri, C. A.; Collet, X.; Fielding, C. J.; Burlingame, A. L. *Protein Sci.* **1995**, 4, 791–803.
- (17) Yoshida, T. *Anal. Chem.* **1997**, 69, 3038–3043.
- (18) Churms, S. C. *J. Chromatogr. A* **1996**, 720, 75–91.
- (19) Alpert, A. J. *J. Chromatogr.* **1990**, 499, 177–196.
- (20) Kieliszewski, M. J.; Oneill, M.; Leykam, J.; Orlando, R. *J. Biol. Chem.* **1995**, 270, 2541–2549.
- (21) Zhang, J.; Wang, D. I. *J. Chromatogr. B* **1998**, 712, 73–82.
- (22) Gobom, J.; Nordhoff, E.; Mirgorodskaya, E.; Ekman, R.; Roepstorff, P. *J. Mass. Spectrom.* **1999**, 34, 105–116.
- (23) Kuster, B.; Mann, M. *Anal. Chem.* **1999**, 71, 1431–1440.
- (24) Gonzalez, J.; Takao, T.; Hori, H.; Besada, V.; Rodriguez, R.; Padron, G.; Shimonishi, Y. *Anal. Biochem.* **1992**, 205, 151–158.
- (25) Xiong, L.; Regnier, F. E. *J. Chromatogr. B* **2002**, 782, 405–418.
- (26) Mills, K.; Johnson, A. W.; Dietrich, O.; Clayton, P. T.; Winchester, B. G. *Tetrahedron-Asymmetr.* **2000**, 11, 75–93.
- (27) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, 68, 850–858.
- (28) Larsen, M. R.; Cordwell, S. J.; Roepstorff, P. *Proteomics* **2002**, 2, 1277–1287.
- (29) Peri, S.; Steen, H.; Pandey, A. *Trends Biochem. Sci.* **2001**, 26, 687–689.
- (30) Viklund, C.; Sjogren, A.; Irgum, K.; Nes, I. *Anal. Chem.* **2001**, 73, 444–452.
- (31) Green, E. D.; Adelt, G.; Baenziger, J. U.; Wilson, S.; Van Halbeek, H. *J. Biol. Chem.* **1988**, 263, 18 253–18 268.
- (32) Harvey, D. J.; Rudd, P. M.; Bateman, R. H.; Bordoli, R. S.; Howes, K.; Hoyes, J. B.; Vickers, R. G. *Org. Mass Spectrom.* **1994**, 29, 753–765.
- (33) Medzihradsky, K. F.; Maltby, D. A.; Hall, S. C.; Settineri, C. A.; Burlingame, A. L. *J. Am. Mass. Spectrom.* **1994**, 5, 350–358.
- (34) Koide, N.; Muramatsu, T. *J. Biol. Chem.* **1974**, 249, 4897–4904.
- (35) Tarentino, A. L.; Plummer, T. H., Jr.; Maley, F. *J. Biol. Chem.* **1974**, 249, 818–824.
- (36) Muramatsu, T. *J. Biol. Chem.* **1971**, 246, 5535–5537.
- (37) Harvey, D. J.; Wing, D. R.; Kuster, B.; Wilson, I. B. *J. Am. Soc. Mass. Spectrom.* **2000**, 11, 564–571.
- (38) Chalkley, R. J.; Burlingame, A. L. *Mol. Cell. Proteomics* **2003**, 2, 182–190.
- (39) Bunkenborg, J.; Pilch, B. J.; Podtelejnikov, A. V.; Wisniewski, J. R. *Proteomics* **2003**, in press.
- (40) Tirumalai, R. S.; Chan, K. C.; Prieto, D. A.; Issaq, H. J.; Conrads, T. P.; Veenstra, T. D. *Mol. Cell. Proteomics* **2003**, 2, 1096–1103.
- (41) Adkins, J. N.; Varum, S. M.; Auberry, K. J.; Moore, R. J.; Angell, N. H.; Smith, R. D.; Springer, D. L.; Pounds, J. G. *Mol. Cell. Proteomics* **2002**, 1, 947–955.
- (42) Pieper, R.; Gatlin, C. L.; Makusky, A. J.; Russo, P. S.; Schatz, C. R.; Miller, S. S.; Su, Q.; McGrath, A. M.; Estock, M. A.; Parmar, P. P.; Zhao, M.; Huang, S. T.; Zhou, J.; Wang, F.; Esquer-Blasco, R.; Anderson, N. L.; Taylor, J.; Steiner, S. *Proteomics* **2003**, 3, 1345–1364.
- (43) Rothemund, D. L.; Locke, V. L.; Liew, A.; Thomas, T. M.; Wasinger, V.; Rylatt, D. B. *Proteomics* **2003**, 3, 279–287.
- (44) Anderson, N. L.; Anderson, N. G. *Mol. Cell. Proteomics* **2002**, 1, 845–864.
- (45) Staudacher, E.; Altmann, F.; Wilson, I. B.; Marz, L. *Biochim. Biophys. Acta* **1999**, 1473, 216–236.
- (46) Trimble, R. B.; Tarentino, A. L. *J. Biol. Chem.* **1991**, 266, 1646–1651.
- (47) Harris, R. J.; Ling, V. T.; Spellman, M. W. *J. Biol. Chem.* **1992**, 267, 5102–5107.

PR034112B