Published in final edited form as:

J Proteome Res. 2016 July 1; 15(7): 2087–2101. doi:10.1021/acs.jproteome.5b00733.

Identification of novel N-glycosylation sites at non-canonical protein consensus motifs

Mark S. Lowenthal^{*}, Kiersta S. Davis, Trina Formolo, Lisa E. Kilpatrick, and Karen W. Phinney

Material Measurement Laboratory, Biomolecular Measurement Division, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8314, Gaithersburg, MD, 20899, USA

Abstract

N-glycosylation of proteins is well known to occur at asparagine residues that fall within the canonical consensus sequence N-X-S/T, but has also been identified at a small number of asparagine residues within N-X-C motifs, including the N₄₉₁ residue of human serotransferrin. Here we report novel glycosylation sites within non-canonical consensus motifs, in the conformation N-X-C, based on mass spectrometry analysis of partially-deglycosylated glycopeptide targets. Alpha-1-acid glycoprotein (A1AG) and serotransferrin (Tf) were observed for the first time to be N-glycosylated on asparagine residues within a total of six unique non-canonical motifs. N-glycosylation was initially predicted *in silico* based on the evolutionary conservation of the N-X-C motif among related mammalian species, and demonstrated experimentally in A1AG from porcine, canine, and feline sources and in human serotransferrin. High-resolution liquid chromatography-tandem mass spectrometry (LC-MS/MS) was employed to collect fragmentation data of predicted GlcNAcylated peptides, and to assign modification sites within N-X-C motifs. A combination of targeted analytical techniques that includes complementary mass spectrometry platforms, enzymatic digestions, and partial-deglycosylation procedures, was developed to confirm the novel observations. Additionally, we found that A1AG in porcine and canine sources is highly N-glycosylated at a non-canonical motif (N-Q-C) based on semi-quantitative multiple-reaction monitoring (MRM) analysis – the first report of an N-X-C motif exhibiting substantial N-glycosylation. Although reports of N-X-C motif N-glycosylation are relatively uncommon in the literature, this work adds to a growing list of glycoproteins reported with glycosylation at various forms of non-canonical motifs.

TOC image

^{*}corresponding author; mark.lowenthal@nist.gov; phone: 301-975-8993.

Introduction

The earliest account of N-glycosylation on a non-canonical motif was reported by Bause and Legler⁶ in 1981 using exogenous synthetic peptides of known amino acid composition as glycosyl acceptors. The *in vitro* work compared catalysis rates of N-glycosyltransferases from calf liver microsomes on synthetic hexa-peptides containing different amino acids at the third position of the consensus motif. The necessity of a hydrogen-bond-acceptor at the third position was established for glycosylation, while the effect of varying the third position amino acid on glycosylation rates was demonstrated to decrease in the order of threonine >> serine > cysteine. The following year, Stenflo and Fernlund were the first to report N-

glycosylation at an N-X-C motif on a natural protein – within the heavy chain of bovine protein C isolated from plasma, and based on Edman degradation of the 260 amino acid-long protein⁷. This finding was later confirmed in bovine plasma protein C and human plasma protein C using immunoassays and Western blotting and led to the hypothesis that N-glycosylation occupancy rates at this non-canonical motif may depend partially on the rate of disulfide bond formation and the rate of protein translation⁸. Later, the primary structure of the 2050 amino acid protein human von Willebrand Factor (vWF) was determined along with a non-canonical N-glycosylation site at N₃₈₄ (N-S-C) through Edman sequencing of the purified and partially proteolyzed protein⁹. Further work mapped the N-glycome of vWF for ten of the canonical motifs (N-X-S/T) and for the N-S-C non-canonical motif using endoglycosidase hydrolysis followed by mass spectrometry detection¹⁰. This study demonstrated nearly all of the vWF canonical consensus sites to be fully occupied; however, the non-canonical N-S-C motif was shown to be only minimally occupied by N-glycosylation. Other proteins that have subsequently been suggested to contain non-canonical N-glycosylation of asparagine residues include human CD69¹¹ (determined by site-directed mutagenesis studies on N₁₁₁ at N-A-C); murine- and fetal- antigen 1^{12, 13} (determined via Edman degradation and a combination of proteolytic and endoglycosidic hydrolysis followed by mass spectrometry analysis); α 1T-glycoprotein¹⁴ isolated from human plasma (glycosylation of N₃₆₂ determined through amino acid analysis with Edman sequencing); human serotransferrin (glycosylation of the N-H-C motif at N₄₉₁ was determined using PNGaseF deglycosylation in H₂¹⁸O and mass spectrometry^{15, 16}); and human factor XI¹⁷ (determined to be $\approx$ 5% occupied at N₁₄₅ (N-I-C) using high-resolution time-of-flight mass spectrometry).

These seven proteins bearing N-glycosylation of non-canonical N-X-C motifs were all derived as natural, purified proteins. However, several studies focused on recombinant protein analysis have also reported N-glycosylation at different non-canonical motifs. The first publication describing recombinant non-consensus N-glycosylation reported the identification of glycan structures at the asparagine within an N-X-G motif found within the C_{H1} constant domain of IgG1 and IgG2 recombinant antibodies¹⁸. The N₁₆₂-S-G atypical motif was modified at estimated levels of 0.5 % to 2.0 %. The same lab later reported an even more atypical N-glycosylation site observed not on an asparagine residue, but on a glutamine residue within a Q-G-T motif, located on recombinant V_L domains of IgG2 molecules¹⁹. The same report also suggests that asparagine may be N-glycosylated in the context of a reverse consensus motif (S/T-X-N). These two reports demonstrate the importance of considering that the expression system or cell type from which a glycoprotein originates will likely have a major effect on the modifications observed. Finally, a very recent study reported low levels of glycosylation for the rare non-canonical motif, N₂₇₄-V-V, on the recombinant protein inter-alpha-trypsin inhibitor heavy chain H4 (ITIH4), and also suggested this modification to be possible in human serum as well²⁰.

Mechanistically, N-glycosylation requires hydrogen bond formation between asparagine and the hydroxy residue (OH or SH) on the C-terminal amino acid in the consensus motif (S/T or C) and with the peptide backbone²¹ in order to facilitate enzymatic transfer of a core oligosaccharide (Glc₃Man₉GlcNAc₂) from the dolichol donor^{22, 23}. Considering that the sulfur on cysteine is less electronegative than the oxygen from serine or threonine (1.8 vs

3.5, Pauling scale), it is not surprising that N-X-C motifs have generally been observed with lower occupancy rates than canonical N-X-S/T motifs, often reported at less than a few percent of the total population of individual glycoproteins. However, it is important to consider that even small changes in the glycoprofile of a protein may have significant physiological or pharmacological effects. Therefore, low level glycosylation has the potential to be both biologically and therapeutically relevant, but this must be determined experimentally on a case-by-case basis. This may be especially significant for understanding biochemical processes and for biopharmaceutical drug development.

This work also considers whether the evolutionary conservation of consensus motifs can be useful *in silico* to predict the occupancy of an N-glycosylation site within non-canonical motifs. Since N-glycosylation must be empirically determined and can be a rather lengthy process, a prediction tool would be a fast, inexpensive, yet valuable screening aid for clinical, industrial, and academic researchers. The proteins considered in this work – serotransferrin and alpha-1-acid glycoprotein – each exhibits considerable sequence conservation at non-canonical N-glycosylation motifs, suggesting a functionally important role for these regions. Predictive algorithms have previously been designed and described in the literature based on the higher-order structure of a glycoprotein in order to calculate the probability that a particular site will be glycosylated^{38, 39}, but these tools require prior knowledge of the protein structure and, more importantly, lack high specificity and sensitivity. N-glycosylation sites have previously been proposed as useful candidates for functional studies⁴⁰, and the use of evolutionary conservation as a tool for identifying functional motifs has been applied to protein phosphorylation across prokaryotic, eukaryotic and mitochondrial proteins⁴¹. Similarly, comparative genomic analyses of N-glycosylation sites have been tested for evolutionary conservation⁴², and N-glycosylation sites have been mapped across evolutionarily distant species⁴³. Yet these reports only considered canonical motifs. We hypothesize that the conservation of non-canonical consensus motifs across species may be a powerful tool to predict whether a site is functionally N-glycosylated.

A common approach for glycan site identification uses the endoglycosidase peptide N-glycosidase F (PNGase F) to cleave N-linked glycans from their peptide backbones. The resulting deamidation reaction that converts the asparagine to aspartic acid can subsequently be detected by mass spectrometry due to a peptide mass shift of + 0.98 Da²⁴⁻³¹. The PNGase reaction is often performed in heavy (¹⁸O) water to amplify the detectable mass shift (+ 2.94 Da) and to distinguish enzymatic deamidation events from any spurious deamidation that may occur during sample preparation. However, this approach is potentially inadequate for low-resolution mass spectrometry detection and also requires careful attention to ensure no complications arise from residual trypsin activity that may catalyze the back-exchange of H₂¹⁸O at the C-terminus of peptides after labeling^{32, 33}, thus resulting in false positives for enzymatic deamidation. As an alternative approach, high mannose, paucimannose, and some hybrid N-glycans can be partially cleaved from their peptide backbones using endo-N-acetylglucosaminidases (Endo H and Endo D). The term “partial deglycosylation” will be henceforth used in this manuscript to describe an enzymatic or chemical cleavage of N-linked glycans at a position other than in between the peptide’s asparagine residue and the reducing terminus GlcNAc monosaccharide. Endoglycosidase enzymes cleave within the chitobiose core, leaving the reducing end

GlcNAc intact on the peptide backbone to be utilized as a + 203.0794 Da “tag” for site identification^{34, 35}. In this study, a targeted analysis is described for N-glycosylation site identification based on a similar enzymatic *partial*-deglycosylation of purified glycoproteins using a cocktail of endoglycosidase F enzymes, followed by targeted LC-MS/MS analysis of the resulting “GlcNAc”ylated peptides. Endo F glycosidases are also specific for N-glycan cleavage at the glycosidic bond within the chitobiose core, but the combinatorial use of Endo F1, Endo F2, and Endo F3 offers broader specificity than a mixture of Endo H and Endo D. By utilizing a mixture of F1, F2, and F3 endoglycosidases, chitobiose core cleavage of high mannose, hybrid, complex bi- and tri-antennary N-glycans can be achieved. This *enzymatic* partial-deglycosylation approach results in the conversion of a fully glycosylated asparagine residue to one which is occupied by a single monosaccharide residue (GlcNAc) and is suitable for glycosite identification using either high resolution or low resolution instrumentation. Orthogonally, a *chemical* partial-deglycosylation approach using trifluoromethane sulfonic acid (TFMS) was used to specifically dehydrate glycosidic bonds, but not amide bonds, resulting in the equivalent + 203.0794 Da GlcNAc tagged-peptide^{36, 37}. Several examples of glycosylation at non-canonical N-X-C motifs are reported in this work, as evidenced through enzymatic or chemical partial-deglycosylation of purified glycoproteins. Occupancy at non-canonical glycosites was predicted based on the evolutionary conservation of the protein’s primary amino acid structure, and subsequently observed using targeted mass spectrometry techniques.

Experimental

Disclaimer—Certain commercial equipment, instruments, and materials are identified in this paper to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by NIST nor does it imply that the equipment, instruments, or materials are necessarily the best available for the purpose.

Materials

Alpha-1 acid glycoprotein purified from serum was purchased for feline (Alpha Diagnostic Int., San Antonio, TX, cat. # A1AG16-N-25), canine (Alpha Diagnostic Int., San Antonio, TX, cat. # A1AG17-N-25), and porcine species (Lee Biosolutions, St. Louis, MO, cat. # 102-14). Human serotransferrin, purified from plasma was purchased from Sigma-Aldrich (St. Louis, MO, cat. # T4382). Endoglycosidases F1, F2, and F3 were purchased from Sigma-Aldrich (F1, cat. # E9762; F2, cat. # E0639); and F3, cat. # E2264). Trifluoromethane sulfonic acid (TFMS) was purchased through Santa Cruz Biotechnology (Dallas, TX). All other chemicals and solvents were acquired from Sigma-Aldrich unless otherwise noted.

Enzymatic digestion

Purified proteins were reconstituted in 100 mmol/L NH_4HCO_3 , pH 7.4 in water and denatured by boiling briefly in 0.2 % (v/v) Rapigest surfactant (Waters, Milford, MA, # 186001861). Approximately 25 μg to 100 μg of protein were reduced with 5 mmol/L (final concentration) dithiothreitol (DTT) (Sigma-Aldrich) by shaking at 60 °C for 30 min, and alkylated with 15 mmol/L (final concentration) iodoacetamide (IAM) (Sigma-Aldrich) in the

dark for 30 min. Additional DTT was used to quench the reaction with IAM. Samples were enzymatically digested using modified porcine trypsin (Promega, Madison, WI, cat. # V5111) at $\approx 1:20$ (enzyme: protein) ratio for 20 h at 37°C with shaking. The reaction was quenched with the addition of HCl to a final concentration of 100 mmol/L. Alternatively, samples were digested using either Lys-C (Promega, V1071) or Glu-C (Promega, # 9PIV165) at similar protein/enzyme ratios. Glu-C digestions were buffered by Tris-HCl, pH 7.2 to avoid interferences with cleavage at Asp. In some cases, sequential digestions were performed using both trypsin and Glu-C. The initial tryptic enzyme reaction was first quenched using HCl, followed by removal of the cleaved Rapigest surfactant via centrifugation at $16000 \times g$ at 4 °C for 10 min. The supernatant containing tryptic peptides was dried to completeness by vacuum centrifugation in a Savant SPD1010 SpeedVac concentrator (Thermo Scientific, Waltham, MA), reconstituted in water, and digested using Glu-C as described above.

Enzymatic partial-deglycosylation

Partial-deglycosylation was performed either prior to, or following, enzymatic digestions. A mixture of endoglycosidases (F1, F2, and F3) was used in a single reaction to cleave N-glycans between the reducing terminal GlcNAc and the remainder of the glycan structure. Enzymes were used at a ratio of U: 0.0048 U (F1), 0.002 U (F2), and 0.003 U (F3) per 100 μ g of protein. The R9150 buffer provided by the manufacturer was used for all of the enzymes within a single buffered reaction. Ten μ L of the R9150 Reaction Buffer was added to HPLC-grade water and to the glycoprotein samples for a total reaction volume of 50 μ L. The reaction was allowed to continue for 20 h at 37° C with constant shaking. The reaction was quenched with the addition of HCl to a final concentration of 100 mmol/L.

Chemical partial-deglycosylation

A chemical partial-deglycosylation of purified glycoproteins or glycopeptides was achieved using trifluoromethane sulfonic acid (TFMS) based on an earlier method³⁷ with minor deviations. Briefly, glycoprotein/glycopeptide samples were dried completely (residual water inhibits this reaction), flushed with N₂, and quickly capped and placed on a dry ice bath for two minutes. Approximately 25 μ L to 50 μ L of toluene was added through an air-tight septum to reconstitute the samples (≈ 20 μ g to 100 μ g total glycoprotein) and the glass vials were kept on a dry ice bath for five min. Neat TFMS (50 μ L) was added, the samples were gently mixed and then placed immediately at -20 °C for four hours. Samples were kept frozen on dry ice for five minutes before quenching the reaction with the addition of 160 μ L of a 3:1:1 (volume ratio) mixture of pyridine: H₂O: MeOH. Samples were left on dry ice for three minutes, moved to -20 °C for five minutes, then to a 4 °C refrigerator for 15 minutes. Ammonium bicarbonate (400 μ L, 50 mmol/L) was added to neutralize the samples prior to glycopeptide enrichment using ZIC-HILIC (SeQuant, Umea, Sweden, Part # 2942-030) or graphitized carbon (Grace #210142, Carbograp SPE Mesh 120/400) SPE cartridges, according to the manufacturer's protocols.

High-resolution targeted LC-MS/MS analysis

A targeted LC-MS/MS analysis for the GlcNAcylated peptides from A1AG and serotransferrin was achieved using a "parent mass inclusion list" for all expected charge

states or dynamic modifications of the monoisotopic peptide mass. Due to the large dynamic range differences in abundance between the most abundant peptides from a typical trypsin digest and those peptides containing partial deglycosylation at non-canonical GlcNAcylation sites, a traditional data-dependent approach would be biased against detection of the glycopeptides of interest. Instead, precursor m/z 's were selected from high-resolution MS¹ scans performed in either a ThermoScientific Orbitrap Elite MS or separately, in an Agilent 6550 quadrupole-time-of-flight (QToF) MS, and fragmented only when specified on a targeted inclusion list.

A ThermoScientific Orbitrap Elite MS was tuned and calibrated using the manufacturer's calibration solution. Peptides were chromatographically separated and eluted in the same way as described above for MRM analyses. FTMS (MS¹) scans were acquired at a resolution of 30000 in positive polarity and profile modes. CID was performed in the linear ion trap and MS² scans were acquired in centroid mode. All CID scans were set as follows: activation Q = 0.250, activation time = 10 ms, normalized collision energy = 35, dynamic exclusion was turned off, data-dependent analysis was performed for the top five most abundant ions from the parent mass inclusion list only as determined from the MS¹ scan, isolation width was set to 2 amu. Source conditions were as follows: heater temperature = 275 °C, sheath gas flow rate = 30 units, auxiliary gas flow rate = 5 units, spray voltage = 3500 V, capillary temperature = 350 °C, and S-lens RF level = 60 %.

An Agilent 6550 QToF MS was tuned and calibrated in standard mode (3200 m/z) in high resolution, extended dynamic range (2 GHz), and was coupled in-line to an Agilent 1260 Infinity HPLC. Peptides were separated on an Atlantis dC₁₈ nanoAcquity (Waters) UPLC column (3 μ m particles, 300 μ m $\times$ 150 mm) at a flow rate of 7 μ L/min. Gradient elution was achieved as described above for MRM analysis. Ionization was performed using a Dual AJS (JetStream) ESI source, and data was collected in profile mode using positive polarity scans. Source conditions were: gas temperature = 275 °C, drying gas = 13 L/min, nebulizer = 343 kPa (35 psig), sheath gas temperature = 350 °C, sheath gas flow = 11 L/min, capillary voltage = 3500 V, nozzle voltage = 1000 V, fragmentor = 1000 V. Five scans/sec or six scans/sec were acquired for MS¹ or MS² scans, respectively. Collision energies were based on the equation $CE = [\text{slope} * (m/z)/100] + \text{offset}$, where slope and offset were set for + 1, + 2, and + 3 ions as 4 and 10, 3.4 and 10, or 2.8 and 10, respectively. Analysis was performed in "Targeted MS/MS" mode with an inclusion list (refer to Table 1 for precursor ions).

Data analysis

All MS² scans were assigned based on *de novo* interpretation. In general, precursor masses for all targeted charge states were extracted from total ion currents and MS² scans were selected for manual interpretation based on known retention times and observed peak intensities. Fragment ions were predicted through software tools (NIST Mass and Fragment Calculator or Skyline) and manually annotated to MS² data. Glycopeptide assignments were made based on the presence of y-ion and b-ion series fragment ions and neutral losses of GlcNAc.

Multiple-reaction monitoring (MRM) LC-MS/MS analysis

Precursor and product ion m/z 's and fragmentation parameters for GlcNAcylated peptides of A1AG and serotransferrin were predicted using the NIST Mass and Fragment Calculator v1.3⁴⁴ and Skyline⁴⁵ (Table 1). The non-GlcNAcylated peptide analogs were monitored as a positive control in each case. Liquid chromatographic separations were achieved using a Zorbax (Agilent) SB-C₁₈ reversed-phase analytical column (2.1 mm × 150 mm, 3.5 μm particles) at a flow rate of 200 μL/min. Peptide separation was achieved using a gradient elution with acetonitrile (ACN) in water up to 50 % (volume ratio) mobile phase B over 35 min, followed by a column wash and re-equilibration. Mobile phases A and B consisted of 0.1 % (volume fraction) formic acid in H₂O or ACN, respectively. Column temperature was maintained at 35 °C. An Agilent 1100 liquid chromatography system was coupled in-line with an ABI 4000 QTrap triple quadrupole (QQQ) mass spectrometer equipped with a standard micro-flow source. Ions were fragmented by collision-induced dissociation (CID) and detected in positive polarity mode with a target scan time of < 3 s and a MRM dwell time > 50 ms. During data acquisition, fragmentation parameters were set to unit resolution in Q1 and Q3, intensity threshold = 0, settling time = 3 ms, pause between mass ranges = 3 ms, ion spray voltage (IS) = 5000 V, capillary temperature = 500 °C, curtain gas = 10 units, GS1 = 40 units, and GS2 = 30 units. Data acquisition was performed using Analyst v1.5 software (Applied Biosystems).

Results

Sequence analysis of glycoproteins

We focused our study on two proteins whose N-glycosylation profiles are well described in the human form and whose amino acid sequences have been elucidated for multiple mammalian species – alpha-1-acid glycoprotein and serotransferrin. After performing sequence alignment across the annotated mammalian species of each glycoprotein using the UniProtKB alignment tool (www.uniprot.org), we manually identified any non-canonical N-X-C (X≠P) motifs and considered the degree of conservation between the genera centering on the hypothesis that more highly conserved sites would be more likely to be N-glycosylated (Figure 1). Selected regions of the protein's primary structure predicted from the genomic data are provided for all manually annotated and reviewed (Swiss-Prot), and non-curated (TrEMBL) datasets. Tryptic peptides that spanned highly conserved motifs were targeted for further analyses as discussed below.

Alpha-1-acid glycoprotein

Alpha-1-acid glycoprotein (A1AG, orosomucoid) is an acute phase plasma protein and one of only four potentially useful circulating biomarkers for all-cause mortality risk⁴⁶. A1AG is also one of the most heavily glycosylated proteins (by mass %) in serum, and as a result is one of the most studied glycoproteins. It contains five well-characterized N-glycosylation sites within canonical N-X-S/T motifs⁴⁷. Our *in silico* analysis revealed a well-conserved, non-canonical motif at N₈₈-X-C in several mammalian species including rabbit (RABIT), cat (FELCA), dog (CANFA), pig (PIG), macaque (MACMU), camel (9CETA), and marmoset (CALJA) (Figure 1a). Interestingly, in nearly all other sequenced mammalian species, the asparagine (N₈₈) residue is mutated to an aspartic acid, leading to the interesting

question of whether this is due to a missense mutation causing the absence/presence of glycosylation and a subsequent loss/gain-of-function, or if it is due to random genetic drift. Inopportunely, in each species of interest, the tryptic peptide that includes N₈₈ within a non-canonical consensus motif also contains a known classical glycosite at N₉₃. The rabbit sequence also includes an asparagine residue at position 87, creating a consensus motif immediately adjacent to the non-consensus motif (N₈₇-N₈₈-T₈₉; N₈₈-T₈₉-C₉₀). Due to the potential for a false positive identification arising from these overlapping motifs, rabbit A1AG was not investigated further.

Purified A1AG was commercially available for three species (porcine, canine, feline) which were considered further for targeted LC-MS/MS analysis. Protein from macaque, camel, and marmoset were unavailable. To reduce the complexity of the glycan compositions we performed partial-deglycosylation of the protein using either an endoglycosidase cocktail (Endo F1, F2, F3) or trifluoromethane sulfonic acid (TFMS) to cleave all but the reducing end N-acetylglucosamine (GlcNAc) residue. Following tryptic digestion, predicted GlcNAcylated glycopeptides were targeted by MRM analysis (Table 1). MRM mass chromatograms are provided in Figure 2 for the +3 charge state of the porcine A1AG tryptic peptide, EYQTIGNQCIYNDSSLK. Glycopeptides were monitored for both singly GlcNAcylated forms (occupancy of either N₈₈ or N₉₃) and the doubly GlcNAcylated peptide (occupancy of both N₈₈ and N₉₃). For each singly occupied peptides, MRM transitions were designed to monitor product ions that are diagnostic for specific site occupancy, whereas the doubly glycosylated peptide was detected by both its unique precursor ion and unique product ions. Figure S1 shows commonalities and differences among the amino acid sequences of the peptide forms. Note that in Figure 2(a,b), both singly glycosylated forms chromatographically co-elute as expected due to the fact that reversed phase separations will be dominated by the peptide backbone, while the doubly glycosylated peptide (Figure 2c) ion elutes only slightly earlier on a C₁₈ phase due to the additional GlcNAc residue.

Six, three, and eight unique transitions were monitored for the EYQTIGNQCIYN*DSSLK, EYQTIGN*QCIYNDSSLK, and EYQTIGN*QCIYN*DSSLK peptides, respectively. The relative MRM peak intensities for the two singly glycosylated forms were observed to be roughly equivalent, suggesting they are glycosylated at similar occupancy rates. An additional GlcNAc residue on the doubly glycosylated form was expected to suppress ionization relative to the singly glycosylated forms. The doubly glycosylated peptide form was observed at a slightly lower signal intensity than either singly glycosylated peptide. Differences in intensity between singly and doubly occupied glycopeptides may be due to variable ionization suppression or fragmentation differences, and/or due to biological differences. Interpretation of this semi-quantitative data should therefore be done with caution.

Identical sample preparations and LC-MS analyses were subsequently performed for the canine and feline protein orthologs. GlcNAcylated peptides were identified at the N₈₈ residues for both species making this the first report of glycosylation at this residue for porcine, canine, and feline A1AG. MRM peak intensities of canine A1AG glycopeptides were observed to be roughly equivalent to those of the porcine protein, while analyses of feline A1AG showed less intense, but detectable glycosylation at N₈₈ (Figure S2). Although

the EYLIGNQCVYNSSFLNVQR tryptic peptide is identical in the canine and feline proteins, the more distantly related porcine A1AG sequence shares only 70 % identity for this segment. This suggests that as the proteins diverged over evolutionary time they mutually retained the N₈₈ glycosylation site.

To confirm the MRM results, the partially-deglycosylated tryptic digests were analyzed by LC-MS in a high-resolution Orbitrap MS followed by MS/MS in the linear ion trap. Figure 3(a–c) provides collision-induced dissociation (CID) MS/MS spectra obtained in the linear ion trap for singly and doubly glycosylated peptides from porcine A1AG, with a complete y-ion series annotated for each glycopeptide to confirm the identification of the novel N₈₈ glycosite in porcine A1AG. Note that as a result of the chromatographic co-elution of the two forms of singly glycosylated glycopeptides and their identical precursor ion mass, their MS/MS fragmentation spectra represent a mixture of the N₈₈-only and N₉₃-only occupied glycopeptides (Fig 3a,b). Shared product ions are assignable to both glycopeptide forms, while the observation of two independent sets of y₆-y₁₀ and b₇-b₁₁ ions that map uniquely to each of the glycopeptide forms confirms the presence of both forms. Detection and fragmentation of the doubly glycosylated peptide along with assignment of a complete y-ion series (Figure 3c) further supports the characterization of the non-canonical glycosite at N₈₈. The most abundant ions in the MS/MS spectrum of the doubly glycosylated peptide were neutral loss ions corresponding to fragmentation of GlcNAc from the peptide. This is to be expected since glycosidic bonds are more labile under CID conditions as compared to the amide peptide bonds.

Serotransferrin

Human serotransferrin contains three non-canonical N-X-C motifs that were targeted for LC-MS analysis, two of which are well conserved among protein analogs from related mammalian species (Figure 2, Table 1). The fully tryptic peptide, INHCR, was observed to be N-glycosylated in this study using two orthogonal, partial-deglycosylation approaches followed by examination of the GlcNAcylated modified peptides. For some samples, deglycosylation was achieved separately following tryptic digestion of the purified glycoprotein using either a cocktail of endoglycosidases (F1, F2, and F3), or through a chemical deglycosylation approach using TFMS. A targeted MRM LC-MS analysis was first performed in a QQQ and subsequently confirmed by fragmentation in an ion trap. Figure 4a demonstrates detection of four independent transitions of the GlcNAcylated IN*HCR peptide from a serotransferrin tryptic digest each identified with signal-to-noise of 9:1. The non-glycosylated peptides were observed by MRM to greater than 100 times the peak intensity of the GlcNAcylated form. However, the weak signal observed during MRM analysis of the glycopeptide does not automatically indicate low biological stoichiometry since glycopeptides are poorly ionized, and CID does not fragment amide bonds as readily as the glycosidic bonds within the glycopeptides. Figure 4b represents the tandem MS/MS spectrum of the modified peptide IN*HCR (* = GlcNAc) that was selected in a Thermo Orbitrap MS, and fragmented in the linear ion trap. The + 2 charge state of the precursor ion (m/z 451.71219) was detected in the orbital trap to within 2.42 mg/kg (ppm) of the theoretical molecular mass, and fragmentation was achieved using collision-induced dissociation (CID). Fragment ions were successfully manually annotated to match a

theoretical spectrum to within 0.8 Da based on the dynamic modification of Asn by a + 203.0794 Da mass shift and static carbamidomethylation of cysteine. Although this is a short peptide, a complete y-ion series was annotated, along with several precursor neutral loss ions. The most abundant fragment ion peaks are due to neutral losses of the GlcNAc residue as discussed for the N₈₈ A1AG glycopeptide. The corresponding Lys-C-digested peptide of serotransferrin, IN*HCRFDEFFSEGCAPGSK, has been previously reported¹⁶ to be N-glycosylated at N₄₉₁ at a rate of < 2 %, and in this study we also observed this longer Lys-C peptide form to be glycosylated using both partial-deglycosylation approaches (endoglycosidases and TFMS).

The fully tryptic peptide from serotransferrin, LCMGSGLNLCEPNNK, was observed to be glycosylated at N₅₂₃ using similar approaches. Figure 5a represents the tandem MS/MS spectrum of the modified form of LCMGSGLN*LCEPNNK after fragmentation by CID in a linear ion trap. The precursor ion was detected in a high-resolution Orbitrap mass spectrometer to within 4 mg/kg (ppm) of the theoretical molecular mass of the GlcNAcylated peptide. As observed with other glycopeptides, the MS/MS spectrum provided a complete y-ion series of fragment ions ensuring high-confidence identification of the modified peptide. Again, the most abundant peaks were observed as neutral losses of GlcNAc. To confirm the presence of glycosylation on N₅₂₃ of serotransferrin, a targeted MRM analysis was performed in parallel. Figure 5b provides mass chromatograms of six fragmentation transitions for the GlcNAcylated form of LCMGSGLN*LCEPNNK. To our knowledge, this non-canonical N-L-C motif has not previously been reported as an occupied glycosylation site.

The third non-canonical glycosylation motif in serotransferrin occurs at N₆₃₇-F-C within the fully tryptic peptide, QQHLFGSNVTDCSGNFCLFR, which also contains the N₆₃₁-V-T canonical motif that has been reported to be highly glycosylated^{48, 49}. The non-canonical N₆₃₇-F-C motif has not previously been reported to be glycosylated, but was observed in this work to be occupied with very low stoichiometry. Using a targeted MRM analysis of the GlcNAcylated tryptic peptide generated via endoglycosidase cocktail treatment, fragmentation transitions containing product ions specific to only one or the other potential N-glycosylation site (N₆₃₁ or N₆₃₇) were selected and monitored on a triple quadrupole mass spectrometer. The doubly GlcNAcylated form of the peptide was concurrently targeted by MRM. This peptide shares some product ions with the singly glycosylated forms, but does not share the same precursor *m/z*. Figure 6a demonstrates the expected glycosylation at the canonical motif with relatively high signal intensity (by MRM analysis) reflecting the reported nearly complete occupancy rate at the N₆₃₁ residue¹⁵. Additionally, Figure 6b shows the mass chromatogram targeting the doubly glycosylated form, QQHLFGSN*VTDCSGN*FCLFR, that, although observed at much lower intensity than the singly glycosylated form, is confidently detected with S/N > 20. Lastly, Figure 6c provides a mass chromatogram specific to glycosylation for the solely occupied non-canonical N₆₃₇-F-C motif. As expected, glycosylation solely at the non-canonical N₆₃₇ site is detected at very low levels due to the fact that lack of glycosylation at the N₆₃₁ residue is very rare. Therefore, it is far less likely for the N₆₃₇ residue to be occupied while the N₆₃₁ is unoccupied, than for both sites to be occupied. It should also be noted that the singly glycosylated forms will sometimes co-elute on a C₁₈ column, further obfuscating detection

of the low abundant, singly occupied N₆₃₇ glycopeptide when targeted for full fragmentation analysis. The doubly glycosylated peptide is less retained than either singly glycosylated peptide (≈ 30 seconds) on a hydrophobic C₁₈ phase.

MS/MS scans of the QQHLFGSN*VTDCSGNFCLFR peptide detected in a QToF mass spectrometer (Figure S3) provide further confirmation of the N₆₃₁ glycosite, but detection of the N₆₃₇ glycosite, which co-elutes, is concealed. Thus it was necessary to perform a second digestion step using Glu-C (selective for cleavage C-terminal to Asp and Glu) following trypsin digestion in order to separate the N₆₃₇ residue from the N₆₃₁ residue. The doubly digested peptide, CSGN*FCLFR, was then targeted for LC-MS/MS analysis. Following the double digestion and partial-deglycosylation using an EndoF cocktail, LC-MS/MS analysis confirmed GlcNAcylation at N₆₃₇ (Figure 7). The precursor ion (682.2921 m/z) was detected in a high-resolution Orbitrap MS within 3 mg/kg (ppm) of the predicted molecular mass. The fragmentation spectrum of the + 2 charge state ion was manually annotated and revealed that the largest peaks were due to neutral losses of the GlcNAc residue. A y-ion series was observed and used to verify peptide identity and the occupancy of the N₆₃₇-F-C motif.

Discussion

Whether considering biological function, development of drug targets, monoclonal antibody characterization, or the search for disease biomarkers, it is essential to consider that even analytes expressed with low stoichiometric abundance may have large biological significance. In general, stoichiometry of non-canonical motif glycosylation should be expected to be quite low, typically less than 2 % occupancy, although the functional impact of the non-canonical glycosylated form may not be directly correlated to abundance. As a result of the low stoichiometry of these modifications, it is challenging for typical data-dependent mass spectrometric analyses (where precursor ion fragmentation is chosen based on abundance) to overcome the challenges of large dynamic range differences between the most and least abundant molecular forms. This is especially relevant when considering that, on a case-by-case basis, glycopeptides may ionize poorly in electrospray mass spectrometry relative to non-glycosylated peptide forms, and that these forms sometimes co-elute. Targeted approaches that specify precursor m/z values for subsequent fragmentation offer a great advantage in this capacity, but this approach could be a daunting task if a protein sequence, or mixture of sequences, contains a vast number of potential glycosylation sites for analysis. The use of conserved N-glycosylation consensus motifs offers a potential avenue for developing targeted analyses focused on sites that are most likely to be occupied.

At the glycopeptide level, several glycoforms often occupy the same glycosite, which creates certain obstacles in the attempt to identify glycosylation sites. First, the presence of multiple glycoforms inherently splits the intensity of the already low-abundant signal and may dampen the signal below the level of detection. Second, for targeted analysis one would have to know or predict the glycan structures themselves in order to calculate the parent peptide mass to identify. We have circumvented these challenges using a partial-deglycosylation strategy, which takes advantage of the fact that all [mammalian] N-glycans share a core glycan structure allowing us to reduce the heterogeneous population of glycan compositions

to the same GlcNAcylated species via enzymatic digest. Little foreknowledge of specific glycan compositions is therefore needed to identify the resulting GlcNAcylated glycopeptides since they all have the same + 203.0794 Da mass tag regardless of their original glycan composition. As long as the peptide sequence is known, the theoretical glycopeptide masses can easily be calculated for targeted analysis. Furthermore, by reducing the multiple glycoforms to one GlcNAcylated form for each occupied site, the total analyte abundance for each glycopeptide is amplified and its detectability is increased. Undeniably this approach does not provide glycan identities, however, this is an important first step to establish novel glycosite identification which can then be followed by characterization of specific glycan compositions. Partial deglycosylation using EndoF or TFMS, allows for the use of lower resolution mass spectrometry analysis, such as by MRM experiments performed on a triple quadrupole MS, introducing the possibility for absolute quantification experiments. In the case where partial-deglycosylation is achieved using EndoF, it is also possible that a core fucosylation could remain on the reducing end GlcNAc residue creating a + 349.1373 Da mass shift. This possibility was included as a dynamic modification for the database searches in the current work, and was observed on several canonical motifs for serotransferrin (data not shown), but has not yet been observed for non-canonical motifs from A1AG or serotransferrin. In the case of partial-deglycosylation using TFMS, however, core fucosylation should not be resistant to the chemical cleavage since the GlcNAc-fucose bond is a C-O glycosidic linkage, and will be cleaved during TFMS hydrolysis.

While this manuscript focuses solely on N-glycosylation of non-canonical N-X-C motifs of A1AG or serotransferrin, the analytical techniques were confirmed based on the known, relatively highly occupied canonical N-X-S/T motifs as positive controls. From observations of spectral counting, GlcNAc'ylated canonical sites were observed significantly more often than GlcNAc'ylated non-canonical sites. In addition to N-glycosylations, the partial-deglycosylation approach using TFMS can be applied towards other types of glycosylation, and has been reported previously for LC-MS analysis of O-glycans^{50, 51}. In the case of partial-deglycosylation of O-glycans, targeted MS analyses must include multiple theoretical mass targets corresponding to all possible reducing end monosaccharides. Although O-glycosylation is not known to rely on a well-characterized consensus "motif", it is specific to R-OH side groups (serine, threonine, and tyrosine) and could be targeted experimentally as such.

The catalytic role of the hydroxy amino acid in the consensus motif has been previously investigated⁶. That work demonstrated that replacing threonine in the consensus motif with serine or cysteine results in a 40- or 100-fold-decrease, respectively, in relative activity of glycosyltransferases. It is also known that a proximal hydrogen-bond acceptor is necessary for the nucleophilicity of asparagine residues to displace the GlcNAc₂Man₉Glc₃ sugar from its dolichol donor, and it is conceivable that this degree of hydrogen-bond accepting potential affects the magnitude of displacement in this co-translational event. N-X-C motifs have been demonstrated here and elsewhere in the literature to be glycosylated in nature, but in nearly all cases are observed at low stoichiometry. This suggests the possibility of a functional role for the S/T/C position in the consensus motif for *regulating* occupancy rates. It has also been suggested that the regulation of glycosyltransferase activity at the consensus

motif may also depend on disulfide bond formation at cysteine residues (protein folding) and the phosphorylation or O-glycosylation at serine or threonine residues⁶.

By itself, the identification of N-linked glycosylation sites is particularly valuable for determining glycoprotein structure and function. However, a more complete understanding of the types of glycan structures at each glycosylation site (high mannose, hybrid, complex) and their decorations (core fucosylation, terminal sialic acids, branching) is desirable. We are currently investigating several approaches for glycan characterization at non-canonical motifs while considering the difficulties associated with the extremely low stoichiometry of these molecules which makes analysis at the fully glycosylated peptide level challenging. The definitive identification of six novel glycosylation sites on non-canonical protein motifs in this work supports and expands upon the current established evidence for N-linked glycosylation on N-X-C motifs. We will continue to exploit the information that can be gained through a reflection on the evolutionary conservation of amino acid motifs towards the prediction and characterization of N-linked glycosylation sites while also applying this approach to other protein modifications suggested to be governed by similar biochemical rules.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Opdenakker G, Rudd PM, Ponting CP, Dwek RA. Concepts and principles of glycobiology. *Faseb J*. 1993; 7(14):1330–7. [PubMed: 8224606]
2. Varki A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*. 1993; 3(2):97–130. [PubMed: 8490246]
3. Bause E, Hettkamp H. Primary structural requirements for N-glycosylation of peptides in rat liver. *FEBS Lett*. 1979; 108(2):341–4. [PubMed: 520572]
4. Marshall RD. The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. *Biochem Soc Symp*. 1974; 40:17–26.
5. Marshall RD. Glycoproteins. *Annu Rev Biochem*. 1972; 41:673–702. [PubMed: 4563441]
6. Bause E, Legler G. The role of the hydroxy amino acid in the triplet sequence Asn-Xaa-Thr(Ser) for the N-glycosylation step during glycoprotein biosynthesis. *Biochem J*. 1981; 195(3):639–44. [PubMed: 7316978]
7. Stenflo J, Fernlund P. Amino acid sequence of the heavy chain of bovine protein C. *J Biol Chem*. 1982; 257(20):12180–90. [PubMed: 6896877]
8. Miletich JP, Broze GJ Jr. Beta protein C is not glycosylated at asparagine 329. The rate of translation may influence the frequency of usage at asparagine-X-cysteine sites. *J Biol Chem*. 1990; 265(19): 11397–404. [PubMed: 1694179]
9. Titani K, Kumar S, Takio K, Ericsson LH, Wade RD, Ashida K, Walsh KA, Chopek MW, Sadler JE, Fujikawa K. Amino acid sequence of human von Willebrand factor. *Biochemistry*. 1986; 25(11): 3171–84. [PubMed: 3524673]
10. Canis K, McKinnon TA, Nowak A, Haslam SM, Panico M, Morris HR, Laffan MA, Dell A. Mapping the N-glycome of human von Willebrand factor. *Biochem J*. 2012; 447(2):217–28. [PubMed: 22849435]
11. Vance BA, Wu W, Ribaudo RK, Segal DM, Kearse KP. Multiple dimeric forms of human CD69 result from differential addition of N-glycans to typical (Asn-X-Ser/Thr) and atypical (Asn-X-cys) glycosylation motifs. *J Biol Chem*. 1997; 272(37):23117–22. [PubMed: 9287313]

12. Krogh TN, Bachmann E, Teisner B, Skjodt K, Hojrup P. Glycosylation analysis and protein structure determination of murine fetal antigen 1 (mFA1)–the circulating gene product of the delta-like protein (dlk), preadipocyte factor 1 (Pref-1) and stromal-cell-derived protein 1 (SCP-1) cDNAs. *Eur J Biochem.* 1997; 244(2):334–42. [PubMed: 9118998]
13. Jensen CH, Krogh TN, Hojrup P, Clausen PP, Skjodt K, Larsson LI, Enghild JJ, Teisner B. Protein structure of fetal antigen 1 (FA1). A novel circulating human epidermal-growth-factor-like protein expressed in neuroendocrine tumors and its relation to the gene products of dlk and pG2. *Eur J Biochem.* 1994; 225(1):83–92. [PubMed: 7925474]
14. Araki T, Haupt H, Hermentin P, Schwick HG, Kimura Y, Schmid K, Torikata T. Preparation and partial structural characterization of alpha 1T-glycoprotein from normal human plasma. *Archives of Biochemistry and Biophysics.* 1998; 351(2):250–256. [PubMed: 9514662]
15. Satomi Y, Shimonishi Y, Hase T, Takao T. Site-specific carbohydrate profiling of human transferrin by nano-flow liquid chromatography/electrospray ionization mass spectrometry. *Rapid Commun Mass Spectrom.* 2004; 18(24):2983–8. [PubMed: 15536627]
16. Satomi Y, Shimonishi Y, Takao T. N-glycosylation at Asn(491) in the Asn-Xaa-Cys motif of human transferrin. *FEBS Lett.* 2004; 576(1–2):51–6. [PubMed: 15474009]
17. Faid V, Denguir N, Chapuis V, Bihoreau N, Chevreux G. Site-specific N-glycosylation analysis of human factor XI: Identification of a noncanonical NXC glycosite. *Proteomics.* 2014; 14(21–22): 2460–70. [PubMed: 25092234]
18. Valliere-Douglass JF, Kodama P, Mujacic M, Brady LJ, Wang W, Wallace A, Yan B, Reddy P, Treuheit MJ, Balland A. Asparagine-linked oligosaccharides present on a non-consensus amino acid sequence in the CH1 domain of human antibodies. *J Biol Chem.* 2009; 284(47):32493–506. [PubMed: 19767389]
19. Valliere-Douglass JF, Eakin CM, Wallace A, Ketchum RR, Wang W, Treuheit MJ, Balland A. Glutamine-linked and non-consensus asparagine-linked oligosaccharides present in human recombinant antibodies define novel protein glycosylation motifs. *J Biol Chem.* 2010; 285(21): 16012–22. [PubMed: 20233717]
20. Chandler KB, Brnakova Z, Sanda M, Wang S, Stalnaker SH, Bridger R, Zhao P, Wells L, Edwards NJ, Goldman R. Site-specific glycan microheterogeneity of inter-alpha-trypsin inhibitor heavy chain H4. *Journal of Proteome Research.* 2014; 13(7):3314–29. [PubMed: 24884609]
21. Imperiali B, Hendrickson TL. Asparagine-linked glycosylation: specificity and function of oligosaccharyl transferase. *Bioorg Med Chem.* 1995; 3(12):1565–78. [PubMed: 8770382]
22. Kornfeld R, Kornfeld S. Assembly of asparagine-linked oligosaccharides. *Annu Rev Biochem.* 1985; 54:631–64. [PubMed: 3896128]
23. Hubbard SC, Ivatt RJ. Synthesis and processing of asparagine-linked oligosaccharides. *Annu Rev Biochem.* 1981; 50:555–83. [PubMed: 7023366]
24. Fan X, She YM, Bagshaw RD, Callahan JW, Schachter H, Mahuran DJ. A method for proteomic identification of membrane-bound proteins containing Asn-linked oligosaccharides. *Anal Biochem.* 2004; 332(1):178–86. [PubMed: 15301963]
25. Bunkenborg J, Pilch BJ, Podtelejnikov AV, Wisniewski JR. Screening for N-glycosylated proteins by liquid chromatography mass spectrometry. *Proteomics.* 2004; 4(2):454–65. [PubMed: 14760718]
26. Qiu R, Regnier FE. Use of multidimensional lectin affinity chromatography in differential glycoproteomics. *Analytical Chemistry.* 2005; 77(9):2802–9. [PubMed: 15859596]
27. Zhang H, Li XJ, Martin DB, Aebersold R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol.* 2003; 21(6):660–6. [PubMed: 12754519]
28. Zhang H, Yi EC, Li XJ, Mallick P, Kelly-Spratt KS, Masselon CD, Camp DG 2nd, Smith RD, Kemp CJ, Aebersold R. High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. *Mol Cell Proteomics.* 2005; 4(2):144–55. [PubMed: 15608340]
29. Liu T, Qian WJ, Gritsenko MA, Camp DG 2nd, Monroe ME, Moore RJ, Smith RD. Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. *Journal of Proteome Research.* 2005; 4(6):2070–80. [PubMed: 16335952]

30. Kaji H, Saito H, Yamauchi Y, Shinkawa T, Taoka M, Hirabayashi J, Kasai K, Takahashi N, Isobe T. Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat Biotechnol.* 2003; 21(6):667–72. [PubMed: 12754521]
31. Morelle W, Donadio S, Ronin C, Michalski JC. Characterization of N-glycans of recombinant human thyrotropin using mass spectrometry. *Rapid Commun Mass Spectrom.* 2006; 20(3):331–45. [PubMed: 16372382]
32. Petritis BO, Qian WJ, Camp DG 2nd, Smith RD. A simple procedure for effective quenching of trypsin activity and prevention of 18O-labeling back-exchange. *Journal of Proteome Research.* 2009; 8(5):2157–63. [PubMed: 19222237]
33. Angel PM, Lim JM, Wells L, Bergmann C, Orlando R. A potential pitfall in 18O-based N-linked glycosylation site mapping. *Rapid Commun Mass Spectrom.* 2007; 21(5):674–82. [PubMed: 17279607]
34. Hagglund P, Bunkenborg J, Elortza F, Jensen ON, Roepstorff P. A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation. *Journal of Proteome Research.* 2004; 3(3):556–66. [PubMed: 15253437]
35. Hagglund P, Matthiesen R, Elortza F, Hojrup P, Roepstorff P, Jensen ON, Bunkenborg J. An enzymatic deglycosylation scheme enabling identification of core fucosylated N-glycans and O-glycosylation site mapping of human plasma proteins. *Journal of Proteome Research.* 2007; 6(8):3021–31. [PubMed: 17636988]
36. Edge AS. Deglycosylation of glycoproteins with trifluoromethanesulphonic acid: elucidation of molecular structure and function. *Biochem J.* 2003; 376(Pt 2):339–50. [PubMed: 12974674]
37. Edge AS, Faltynek CR, Hof L, Reichert LE Jr, Weber P. Deglycosylation of glycoproteins by trifluoromethanesulfonic acid. *Anal Biochem.* 1981; 118(1):131–7. [PubMed: 6175244]
38. Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, Meyer AG, Wilke CO. Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. *J Mol Evol.* 2014; 79(3–4):130–42. [PubMed: 25217382]
39. Lam PV, Goldman R, Karagiannis K, Narsule T, Simonyan V, Soika V, Mazumder R. Structure-based comparative analysis and prediction of N-linked glycosylation sites in evolutionarily distant eukaryotes. *Genomics Proteomics Bioinformatics.* 2013; 11(2):96–104. [PubMed: 23459159]
40. Kim DS, Hahn Y. The acquisition of novel N-glycosylation sites in conserved proteins during human evolution. *BMC Bioinformatics.* 2015; 16(29):015–0468.
41. Gnäd F, Forner F, Zielinska DF, Birney E, Gunawardena J, Mann M. Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria. *Mol Cell Proteomics.* 2010; 9(12):2642–53. [PubMed: 20688971]
42. Park C, Zhang J. Genome-wide evolutionary conservation of N-glycosylation sites. *Mol Biol Evol.* 2011; 28(8):2351–7. [PubMed: 21355035]
43. Zielinska DF, Gnäd F, Schropp K, Wisniewski JR, Mann M. Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. *Mol Cell.* 2012; 46(4):542–8. [PubMed: 22633491]
44. Kilpatrick EL, Liao WL, Camara JE, Turko IV, Bunk DM. Expression and characterization of 15N-labeled human C-reactive protein in *Escherichia coli* and *Pichia pastoris* for use in isotope-dilution mass spectrometry. *Protein Expr Purif.* 2012; 85(1):94–9. [PubMed: 22796447]
45. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics.* 2010; 26(7):966–968. [PubMed: 20147306]
46. Fischer K, Kettunen J, Wurtz P, Haller T, Havulinna AS, Kangas AJ, Soininen P, Esko T, Tammesoo ML, Magi R, Smit S, Palotie A, Ripatti S, Salomaa V, Ala-Korpela M, Perola M, Metspalu A. Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons. *PLoS Med.* 2014; 11(2)
47. Treuheit MJ, Costello CE, Halsall HB. Analysis of the five glycosylation sites of human alpha 1-acid glycoprotein. *Biochem J.* 1992; 283(Pt 1):105–12. [PubMed: 1567356]

48. Spik G, Debruyne V, Montreuil J, van Halbeek H, Vliegthart JF. Primary structure of two sialylated triantennary glycans from human serotransferrin. *FEBS Lett.* 1985; 183(1):65–9. [PubMed: 3979568]
49. MacGillivray RT, Mendez E, Sinha SK, Sutton MR, Lineback-Zins J, Brew K. The complete amino acid sequence of human serum transferrin. *Proc Natl Acad Sci U S A.* 1982; 79(8):2504–8. [PubMed: 6953407]
50. Gerken TA, Owens CL, Pasumathy M. Determination of the site-specific O-glycosylation pattern of the porcine submaxillary mucin tandem repeat glycopeptide. Model proposed for the polypeptide:galnac transferase peptide binding site. *J Biol Chem.* 1997; 272(15):9709–19. [PubMed: 9092502]
51. Muller S, Goletz S, Packer N, Gooley A, Lawson AM, Hanisch FG. Localization of O-glycosylation sites on glycopeptide fragments from lactation-associated MUC1. All putative sites within the tandem repeat are glycosylation targets in vivo. *J Biol Chem.* 1997; 272(40):24780–93. [PubMed: 9312074]

a) alpha-1-acid glycoprotein

```

SP|P02763|A1AG1_HUMAN  QEIQATFFYFTPNKTEDTIFLREYQTRQDQCIYNTTYLNVQRENGTISRYYVGGQEHFAHL 119
SP|P02764|A1AG_RAT     QTIQTEYFYLTPLNDTIELREFQTTDDQCVYNFTHLGVQRENGTSLKAGAVKIFFAHL 120
SP|P25227|A1AG_RABIT   QHTQAAFFYFTAIEEDTLLREYITNTGTFYNSSIVRVQRENGTSLKHDGIRNSVADL 119
SP|Q60590|A1AG1_MOUSE  QTMQSEFFYLTTNLNDTIELRESQTIGDQCVYNSTHLGFQRENGTFSKYEGGVETFAHL 120
SP|P21350|A1AG1_MUSCR  QKMQMVFVNITPNLNDTIELREYHTIDDDHCVYNSTHLGIQRENGTSLKYVGGVKIFADL 120
SP|Q3S2R3|A1AG_BOVIN   RAIQAAFFYLEPRHAEDKLITREYQTIEDKCVYNCSFKIYRQNGTSLKVESDREHFVDL 120
TR|M3VU75|M3VU75_FELCA RTIQAFFYFTHINYTEDKILLREYLTIGNOCVYNSSYLNVQRENGTSLKHEFGKEQVGYL 119
TR|G1M1M0|G1M1M0_AILME RTIHAAFFYFAPNHTDDKILLREYQTIQDQCVYNSSYLKQVRENGTSLKYEYEGKEQFADL 119
TR|G1QF01|G1QF01_MYOLU SKIHAVFFYFTPNHTEDTILLREYITVGDQCIYNTSSVKVQRENGTSLKSENGTEHFVHL 118
TR|F6Y713|F6Y713_CANFA RTIHAAFFYFDVNHTDDTILLREYLTIGNOCVYNSSFLNVQRENGTSLKYEYEGKENFGIL 119
TR|F7H7G9|F7H7G9_MACMU HEIQATFFYFTPNKTEDTIFLREYQTRQDQCIYNTSYLNVQRENGTISKYEYEGQEHFAHL 119
TR|A7Y9J5|A7Y9J5_CAPHI RAIQAAFFYFEPHRAEDKCMAREYQTIADKCVYNCSYINVYRQNGTSLKFETDREHFADL 120
TR|L9LA59|L9LA59_TUPCH RTIHANFFYFTPNHTEDQILLREYMTVGDQCIYNSDLVQVQVNGTISKHEWGREHFAHL 119
TR|F1SN68|F1SN68_PIG   RSIQAFFFFDPKPAEDKINLREYQTIQDQCIYNTSSSLKVVHRENGTSLKHEMGREHVAHL 119
TR|G1S4I5|G1S4I5_NOMLE HEIQATFFYFTPNKTEDTIFLREYQTRQDQCFYNTTYLNVQRENGTISRVEGGREHFAHL 119
TR|G3HHZ1|G3HHZ1_CRIGR QKVQAGYFYFTPNLTDITLLQEQYQTKEDRCVYNSSKLGVQRENGTISKYEGLVEHVAHL 119
TR|G3QME1|G3QME1_GORGO QEIQATFFYFTPNKTEDTIFLREYQTRQDQCFYNTTYLNVQRENGTISRVEGGREHAAHL 119
TR|W5P7S6|W5P7S6_SHEEP RAIQAAFFYFEPHRAEDKLIVREYQTIADKCVYNCSYINVYRQNGTSLKLEMDREHFADL 120
TR|H0X1W6|H0X1W6_OTOGA MEISADFFYLAPNKTEDKIELREYQTIQDQCVYNFSYLNVQRENGTISKYEYEGGKEHYAHL 119
TR|G3TBT9|G3TBT9_LOXAF GEIQATFFYFTPNITEDVILLRQYTTMRGQCIYNSSYLGIQRENGTSLKVFQNGFVNL 119
TR|I3MJR5|I3MJR5_SPETR QKVQASDFSFTPNLTQDITLQYQYLSLEDHCEYNSSVLVWHRENATLSRFDDGGKKIDAHL 119
TR|S9WX48|S9WX48_9CETA RSIQAFFYFSPNYVEDKILLREYMTIGNOCVYNSSYILTVYRENGTISKHEWGREHFAHL 119
TR|B6D983|B6D983_CAPIB RAIQAAFFYFEPHRAEDKLIVREYQTIADKCVYNCSYINIYRQNGTSLKFETDREHFADL 120
TR|F7ISF2|F7ISF2_CALJA DIQ-ATFFYFTPNKTEDTIFLREYQTRNACVYNSSYLNVQRENGTISKFEEGREHVAHL 119

```

* : * : : * * : : . * * : : *

Homo sapiens (Human); Rattus norvegicus (Rat); Oryctolagus cuniculus (Rabbit); Mus musculus (Mouse); Mus caroli (Ricefield mouse); Bos taurus (Bovine); Felis catus (Cat); Ailuropoda melanoleuca (Giant panda); Myotis lucifugus (Little brown bat); Canis familiaris (Dog); Macaca mulatta (Rhesus macaque); Capra hircus (Goat); Tupaia chinensis (Chinese tree shrew); Sus scrofa (Pig); Nomascus leucogenys (Northern white-cheeked gibbon); Cricetulus griseus (Chinese hamster); Gorilla gorilla gorilla (Western lowland gorilla); Ovis aries (Sheep); Otlemur garnettii (Greater bushbaby); Loxodonta Africana (African elephant); Spermophilus tridecemlineatus (Ground squirrel); Camelus ferus (Wild Bactrian camel); Capra ibex (Ibex); Callithrix jacchus (White-tufted-ear marmoset)

```

TR|Q5R9L7|Q5R9L7_PONAB      KDKEDCVKVLNQQ--QRLFGSNVT----DCSGNFCLFRSETKDLLFRDDTVCLAKLHNR 662
TR|M3WBQ5|M3WBQ5_FELCA      KDMAACVSKTLLDQ--QKLFGKT----ENDCSSQFCMFQSETKDLLFKDDTKCLAKLPEG 673
TR|A0A0A7RU99|A0A0A7RU99_SAISC KDKAVCVQKMLFDQ--QKIFGSSVS----DCSSYFCMFRSETKDLLFRDDTECLAKLHDK 665
TR|A0A0A7RNI8|A0A0A7RNI8_PAPAN KDKADCVQTLNQQ--QRMFGSSVTT--PNNCSSNFCLFESKTKDLLFRDDTVCLAKLHDK 666
TR|A0A0A7RQ70|A0A0A7RQ70_AOTNA KDKADCVRQMLNQQ--QKIFGRSVP----DCSSYFCMFRSETKDLLFRDDTVCLAKLYDK 664
TR|A0A0A7RNG5|A0A0A7RNG5_HYLSY KDKEDCVKVLNQQ--QRIFGSNVT----DCSGNFCLFRSETKDLLFRDDTVCLAKLHNR 663
TR|A0A0A7RNJ2|A0A0A7RNJ2_TRAFR KDKADCVHTLLDQ--QRMFGSSVT----DCSSNFCLFRSETKDLLFRDDTVCLAKLHNR 665
TR|A0A0A7RNC8|A0A0A7RNC8_NOMLE KDKEDCVKVLNQQ--QRIFGSNVT----DCSGNFCLFRSETKDLLFRDDTVCLAKLHNR 663
TR|B8R1K3|B8R1K3_BOSMU      KDKATCVKILLNQQ--QDDFGKS----VTDCTSNFCLQSNKSKDLLFRDDTVCLASI-AK 669
TR|G3R4X1|G3R4X1_GORGO      KDKACVHKILLNQQ--QHLFGSNVT----DCSGNFCLFRSETKDLLFRDDTVCLAKLHNR 663
TR|D2HNS5|D2HNS5_AILME      KDKASCVKILLLEQ--QLLFGGS----GNDCSGKFCLFHSSETKDLLFRDDTVCLAKLPDG 662
TR|G7NXY9|G7NXY9_MACFA      KDKADCVQTLNQQ--QRMFGSSVT----DCSSNFCLFESKTKDLLFRDDTVCLAKLHNR 663
TR|A0A0A7RND2|A0A0A7RND2_ALLNI KDKADCVQTLNQQ--QRMFGSSVT----DCSSNFCLFESKTKDLLFRDDTVCLAKLHNR 663
TR|A0A0A7RU88|A0A0A7RU88_HYLLA KDKEDCVKVLNQQ--QRLFGSNVT----DCSGNFCLFRSETKDLLFRDDTVCLAKLHNR 663
TR|A0A0A7RU93|A0A0A7RU93_CERAS KDKADCVQTLNQQ--QRMFGSSVT----DCSSNFCLFESKTKDLLFRDDTVCLAKLHNR 663
TR|Q7TSX8|Q7TSX8_MARMO      KDKAACVREVLRSQ--VTEFGSH----VSDCSNKFCLFSSETKDLLFRDDTVCLVRLTDD 659
TR|G7MJR4|G7MJR4_MACMU      KDKADCVQTLNQQ--QRMFGSSVT----DCSSNFCLFESKTKDLLFRDDTVCLAKLHNR 663
TR|F7HZZ3|F7HZZ3_CALJA      KDKADCVQVLLDQ--QKIFGRSVP----DCSSYFCMFRSETKDLLFRDDTVCLAKLHDK 670
TR|M3YE64|M3YE64_MUSPF      KDKASCVKILLLEQ--QTMFGGN----GNDCSGKFCLFHSSETKDLLFRDDTVCLAKLPES 661
TR|W5PF65|W5PF65_SHEEP      KDKATCVIRILLKEQ--QANFGKA----VTDCTSNFCLQSTSKDLLFRDDTVCLASI-AK 669
TR|H0XSQ4|H0XSQ4_OTOGA      KDKAACVRQTLRHQENLRWPG-----EADCSKKFCMFQSDTKDLLFKDNTCLAQIQDG 670
TR|I3MMA5|I3MMA5_SPETR      KDKAACVREVLRSQ--VMKYGSH----VSDCSSEFCFLSSETKDLLFRDDTVCLVRLPDD 666
TR|J9P430|J9P430_CANFA      KDKASCVKMLLDQ--QLLFGRN----GNDCSGKFCLFHSATKDLLFRDDTVCLAKLPED 721
.: * * * * *: : : :

```

Homo sapiens (Human); Rattus norvegicus (Rat); Mus musculus (Mouse); Sus scrofa (Pig);
 Oryctolagus cuniculus (Rabbit); Bos taurus (Bovine); Equus caballus (Horse); Pan troglodytes
 (Chimpanzee); Pteropus alecto (Black flying fox); Heterocephalus glaber (Naked mole rat); Myotis
 davidii (David's myotis); Cricetulus griseus (Chinese hamster); Myotis brandtii (Brandt's bat);
 Pongo pygmaeus pygmaeus (Bornean orangutan); Callithrix geoffroyi (Geoffroy's marmoset); Colobus
 guereza (Mantled guereza); Alouatta sara (Bolivian red howler monkey); Lagothrix lagotricha
 (Brown woolly monkey); Callicebus moloch (Dusky titi monkey); Saguinus fuscicollis (Brown-headed
 tamarin); Pongo abelii (Sumatran orangutan); Felis catus (Cat); Saimiri sciureus (Common squirrel
 monkey); Papio anubis (Olive baboon); Aotus nancymae (Ma's night monkey); Hylobates syndactylus
 (Siamang); Trachypithecus francoisi (Francois' leaf monkey); Nomascus leucogenys (Northern white-
 cheeked gibbon); Bos mutus grunniens (Wild yak); Gorilla gorilla gorilla (Western lowland
 gorilla); Ailuropoda melanoleuca (Giant panda); Macaca fascicularis (Crab-eating macaque);
 Allenopithecus nigroviridis (Allen's swamp monkey); Hylobates lar (Common gibbon); Cercopithecus
 ascanius (Black-cheeked white-nosed monkey); Marmota monax (Woodchuck); Macaca mulatta (Rhesus
 macaque); Callithrix jacchus (White-tufted-ear marmoset); Mustela putorius furo (European
 domestic ferret); Ovis aries (Sheep); Oryzomys latipes (Small-eared galago); Sperophilus
 tridecemlineatus (Thirteen-lined ground squirrel); Canis familiaris (Dog)

Figure 1.

Alignment of primary amino acid sequences from related species (mammalian; www.uniprot.org; UniProtKB release 2015-05) for selected regions of each protein a) alpha-1-acid-glycoprotein, and b) serotransferrin containing non-canonical N-X-C motifs (highlighted in green), and known, potentially interfering N-glycosylation sites (canonical motifs) (highlighted in yellow). Tryptic peptides targeted by LC-MS are shown in red rectangles. A phylogenetic tree inferring evolutionary relationship between species is also provided for each protein.

Figure 2.

Mass chromatograms of the tryptic peptides, a) EYQTIGN*QcIYNDSLLK, b) EYQTIGNQcIYN*DSSLK, and c) EYQTIGN*QcIYN*DSSLK targeted by LC-MS/MS (MRM) analysis on a triple quadrupole (QQQ) system. Glycopeptides from porcine alpha-1-acid glycoprotein (A1AG) were subjected to partial-deglycosylation prior to MRM analysis. Both the non-canonical (NQC) and canonical (NDS) motifs are shown to be glycosylated, as is the doubly glycosylated form of the protein. n* indicates GlcNAcylated Asn; c* indicates carbamidomethylated Cys.

Figure 3.

Tandem mass spectra (MS²) of the tryptic peptides, a) EYQTIGNQcIYN*DSSLK, b) EYQTIGN*QcIYNDSSLK, and c) EYQTIGN*QcIYN*DSSLK targeted by LC-MS/MS analysis of porcine alpha-1-acid glycoprotein (A1AG). Glycopeptides were subjected to partial-deglycosylation prior to LC-MS analysis. Precursor ions were detected in a high-resolution Orbitrap Elite and MS² spectra were collected in the linear ion trap. Glycosylation of both the non-canonical (NQC) and canonical (NDS) motifs are detected from a shared MS² spectrum, while the doubly glycosylated form of the protein is determined independently. n* indicates GlcNAcylated Asn; c* indicates carbamidomethylated Cys.

Figure 4.
a) MRM mass chromatogram, and b) tandem mass spectrum (MS²) of the glycopeptide In*HcR (* = GlcNAc) observed from the targeted analysis of the fully tryptic peptide from human serotransferrin. Glycopeptides were subjected to partial-deglycosylation prior to LC-MS analysis. The Lys-C digested form of this glycosylated peptide (In*HcRFDEFFSEGcAPGSK) has been reported previously, and also confirmed by this work. n* indicates GlcNAcylated Asn; c* indicates carbamidomethylated Cys.

Figure 5.

a) Tandem mass spectrum (MS^2), and b) MRM mass chromatogram of the glycopeptide LcMGSGLn*LcEPNNK (* = GlcNAc) observed from the targeted analysis of the fully tryptic peptide from human serotransferrin. Glycopeptides were subjected to partial-deglycosylation prior to LC-MS analysis. The glycosylated form of this peptide has not been reported previously. n* indicates GlcNAcylated Asn; c* indicates carbamidomethylated Cys.

Figure 6.

LC-MS/MS (MRM) mass chromatograms from the targeted analysis of three forms of the variably glycosylated peptide QQHFLFGSNVTDCSGNFCLFR observed from human serotransferrin. Glycopeptides were subjected to partial-deglycosylation prior to LC-MS analysis. Occupancy of a) only the canonical motif, QQHFLFGSn*VTDcSGNFcLFR, was observed with robust signal; b) both glycosites, QQHFLFGSn*VTDcSGn*FcLFR, was also observed with very good S/N (> 20:1); and c) only the non-canonical motif, QQHFLFGSNVTDCSGn*FcLFR, was not observed to have clear signal above noise, as expected from the known near-100% occupancy rate of the canonical motif. n* indicates GlcNAcylated Asn; c* indicates carbamidomethylated Cys.

Figure 7.

Tandem mass spectrum (MS^2) of the glycopeptide cSGn*FcLFR (* = GlcNAc) observed from the targeted analysis of the double digestion (trypsin and Glu-C) of human serotransferrin. The glycopeptide was subjected to partial-deglycosylation using Endo F prior to LC-MS analysis. The glycosylated form of this peptide has not been reported previously. n* indicates GlcNAcylated Asn; c* indicates carbamidomethylated Cys.

Table 1

Targeted peptides, theoretical masses, and experimental m/z 's with MRM fragmentation parameters used for the analysis of N-linked glycosylation on A1AG and serotransferrin. Theoretical parent masses with an additional + 146.06 Da were also included when searching glycopeptide data generated by enzymatic partial-deglycosylation to account for potential core fucosylation; since no such species were identified at non-canonical sites these data are not included in the table. n* indicates GlcNAcylated Asn; c* indicates carbamidomethylated Cys.

Glycoprotein	Species	Digestion Enzyme(s)	Peptide	Theoretical Mass (Da)	Precursor ion <i>m/z</i> (Q1), Da	Product ion <i>m/z</i> (Q3), Da	Precursor charge state	Fragment Ion	DP (volts)	CE (volts)
Alpha-1-acid glycoprotein	Porcine	Trypsin	EYQTIGNQcLYn*DSSLK	2235.0001	746.007	MRM transitions	+3	+3y9	85.5	30.2
						1302.6				
						1142.6				
						1029.5				
						651.8				
						571.8				
						515.2				
						826.4				
			663.3							
			413.7							
			EYQTIGn*QcLYNDSSLK			1142.6				
						1029.5				
						434.3				
						347.2				
						902.9				
						874.4				
						571.8				
	515.2									
	EYQTIGn*QcLYn*DSSLK	1326.7								
		1227.6								
		1064.5								
1529.8										
Canine/Feline		EYLTIGn*QcVYnSSFLNVQR	2607.2275	1304.621	+2	+2y11	126.2	70.1		
						+2y10				
						+2y9				
						+2y11				

Glycoprotein	Species	Digestion Enzyme(s)	Peptide	Theoretical Mass (Da)	Precursor ion <i>m/z</i> (Q1), Da	Product ion <i>m/z</i> (Q3), Da	Precursor charge state	Fragment Ion	DP (volts)	CE (volts)
Transferrin	Human	Trypsin	EYLTIGn*QcVYn*SSFLNVQR	2810.3069	1406.161	1430.7	+2	+2y10	133.6	75.9
						1267.6		+2y9		
						1430.7		+2y10		
						1267.6		+2y9		
						776.4		+2y6		
			629.4	+2y5						
			In*HcR	901.4076	902.415	+1	+1y4	96.9	47.2	
							472.2			+1y3
							335.2			+1y2
							568.3			+1b3
no MRM data										
Transferrin	Human	Trypsin	LcMGSGLn*LcEPNNK	1908.8380	955.426	1361.6	+2	+2y10	100.8	50.2
						1304.6		+2y9		
						1191.5		+2y8		
						681.3		+2y10+2		
						652.8		+2y9+2		
			596.3	+2y8+2						
			QQQHFLFGSn*VTDCSGn*FcLFR	2920.2756	974.432	+3	+3y16+2	102.2	37.3	
							1070.5			+3y15+2
							1041.9			+3y14+2
							682.3			+3y9+2
602.3	+3y8+2									
QQQHFLFGSn*VTDCSGNFCfLFR	2717.1962	906.739	+3	+3y7+2	97.2	35.2				
				558.8			+3b11+2			
				722.3			+3b12+2			
				779.9			+3b13+2			
				859.9			+3b13+2			
In*HcRDEFFSEGCAPGSK	2460.0474	1231.031	+2	+2b8	120.9	65.9				
				1275.5			+2b9			
				1422.6						

Glycoprotein	Species	Digestion Enzyme(s)	Peptide	Theoretical Mass (Da)	Precursor ion <i>m/z</i> (Q1), Da	Product ion <i>m/z</i> (Q3), Da	Precursor charge state	Fragment Ion	DP (volts)	CE (volts)				
						1569.7		+2b10						
					821.023	638.3	+3	+3b8+2	91.0	32.5				
						711.8		+3b9+2						
						785.3		+3b10+2						
									682.290	no MRM data				
									1362.5697					
					Trypsin & Glu-C				cSGn*FCLFR					