Published in final edited form as:

J Proteome Res. 2007 August ; 6(8): 3162–3173. doi:10.1021/pr070111y.

N-Glycan Structure Annotation of Glycopeptides Using a Linearized Glycan Structure Database (GlyDB)

Jian Min Ren[§], Tomas Rejtar, Lingyun Li, and Barry L. Karger^{*}

Barnett Institute and Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115

Abstract

While glycoproteins are abundant in nature, and changes in glycosylation occur in cancer and other diseases, glycoprotein characterization remains a challenge due to the structural complexity of the biopolymers. This paper presents a general strategy, termed GlyDB, for glycan structure annotation of N-linked glycopeptides from tandem mass spectra in the LC-MS analysis of proteolytic digests of glycoproteins. The GlyDB approach takes advantage of low-energy collision induced dissociation of N-linked glycopeptides that preferentially cleaves the glycosidic bonds while the peptide backbone remains intact. A theoretical glycan structure database derived from biosynthetic rules for N-linked glycans was constructed employing a novel representation of branched glycan structures consisting of multiple linear sequences. The commonly used peptide identification program, Sequest, could then be utilized to assign experimental tandem mass spectra to individual glycoforms. Analysis of synthetic glycopeptides and well-characterized glycoproteins demonstrate that the GlyDB approach can be a useful tool for annotation of glycan structures and for selection of a limited number of potential glycan structure candidates for targeted validation.

Introduction

Glycosylation, the most extensive and complex form of protein post-translational modification (PTM),^{1, 2} significantly affects cellular functions, including cell apoptosis, angiogenesis, anticoagulation, cell-cell communication and microbial pathogenesis.^{1, 3-5} Most of the key molecules involved in the innate and adaptive immune response are glycoproteins.⁶ Her2⁷, PSA^{8, 9} and CA125¹⁰ are examples of glycoproteins currently used in clinical practice as diagnostic cancer biomarkers. Moreover, glycosylation changes in cancer¹¹ and other diseases¹² greatly impact such processes as metastasis.¹³ However, despite the significance of glycoproteins, only a small percentage of these substances have been fully characterized.²

Glycosylation is post-translationally attached to a protein in a non-template driven process, and thus, compared to protein sequences, the structure cannot be directly predicted. In addition, each glycosylation site contains a variety of glycan structures, leading to a highly complex mixture. In order to determine various glycoforms of the same glycoprotein, glycans are typically released from the glycoproteins, either enzymatically or chemically, and then separated using, for example, ion exchange chromatography, reversed phase chromatography, or capillary electrophoresis, either in the native state¹⁴ or after derivatization.¹⁵

Recently, various mass spectrometric (MS) approaches have been developed for systematic analysis of glycan structures.¹⁶⁻²⁷ Typically, masses of released glycans are determined by

* Corresponding author, email: b.karger@neu.edu.

[§]Current address: Cell Signaling Technology, Inc., Danvers, MA 01923

MS analysis, and the most likely glycan compositions are then assigned to individual glycoforms.¹⁹ However, this approach provides only limited structural information. Other strategies rely on the determination of glycan structure based on MS/MS fragmentation,^{21, 28} allowing determination of not only glycan composition but also partial linkage information, and in some cases, even discrimination between isomeric monosaccharides.²⁹ The fragmentation of various classes of glycans has been extensively studied,³⁰⁻³³ leading to development of computer algorithms that can assist the interpretation of glycan MS/MS spectra. As in the case of peptide identifications, such programs utilize a database of glycan structures¹⁶, a library of experimental spectra³⁴ or the algorithms result in *de novo* spectral interpretation.^{22, 35-37} Recently, a software package that allows comprehensive elucidation of glycan structures using MSⁿ spectra of permethylated glycans has been introduced.^{21, 34, 38} Despite the progress in interpretation of the mass spectra of glycans, the data acquisition as well as data interpretation still require manual intervention.

Although the analysis of released glycans facilitates high throughput analysis and comparison of glycosylation patterns between samples, the approach provides only an overview of glycosylation, since the information on association between glycans and proteins is generally lost. In addition, large amounts of sample are typically required because multiple sample preparation steps can lead to losses,^{16-18, 25-27, 39} and glycans are not very sensitive, unless derivatized, in terms of electrospray or MALDI ionization.

Compared to analysis of glycans, less attention has been paid to the direct analysis of glycopeptides despite the fact that the latter can be easily generated and separated by HPLC without the need for derivatization. The attached peptide can be viewed as a derivatization tag since the peptide moiety can significantly increase ionization efficiency, relative to the glycans themselves. Moreover, the glycoprotein identity, including the location of the glycosylation site, even in the case of a mixture of glycoproteins, can be determined, for example, by electron capture (ECD)⁴⁰ or electron transfer dissociation (ETD), which, unlike low energy CID, leads preferentially to cleavage of the peptide backbone.⁴¹

Low energy tandem CID mass spectra of glycopeptides typically contain fragment ions corresponding to cleavage along the glycosidic bonds with little or no peptide fragmentation.⁴² Importantly, the occurrence of only glycosidic fragments of glycopeptides allows relatively simple interpretation of MS/MS spectra. On the other hand, the lack of cross-ring fragments limits determination of the exact linkage between monosaccharide units, as well as the identity of individual monosaccharides. If necessary, characterization of linkages can be performed separately using other techniques, e.g. digestion of glycans with specific enzymes or some specific glycosylation patterns, e.g. glycans and their linkages at the core, are known based on the given species.⁴³ Nevertheless, interpretation of low energy CID spectra of glycopeptides has the potential of high throughput trace characterization of glycoproteins.

This paper presents a simple approach, GlyDB, for the annotation of tandem mass spectra of N-linked glycopeptides using a custom-built linearized glycan structure database and the well-developed search engine, Sequest. The GlyDB approach relies on a theoretical database of potential glycan structures derived from glycan biosynthetic rules for annotation of glycopeptide tandem mass spectra. The GlyDB approach is shown to be particularly useful for initial annotation of MS/MS spectra with assignment of potential glycan structures. The structures can then be manually validated. Importantly, the use of glycopeptides, rather than glycans, removes the deglycosylation and glycan derivatization steps, leading in general to more sensitive analyses. The examples presented in this paper demonstrate the usefulness of the approach to glycoprotein structural analysis.

GlyDB Overview

As noted above, the GlyDB approach was developed to address the need for annotation of glycopeptide structures in the LC-ESI MS analysis of proteolytic digests of glycoproteins. The approach is based on a similar strategy to the identification of peptides using database searching, and thus, for purposes of clarity, the principles of peptide identification will first be briefly reviewed.

Most of the current peptide identification strategies rely on the ability of a database search algorithm to assign an experimental MS/MS spectrum to the most likely peptide out of all sequences potentially present in the sample. The possible peptides are generated using an *in-silico* proteolytic digestion of a specific database of proteins. In order to limit the number of potential matches, only peptides with m/z values within the experimental precursor ion mass tolerance are considered. Next, the database search algorithm calculates a similarity score between the theoretically generated MS/MS spectra of all potential peptides and the experimental MS/MS spectrum. Matches with similarity scores greater than a specified threshold are typically accepted as significant identifications.

The GlyDB approach utilizes the same principles as peptide identification to assign the most likely glycan structure to a given MS/MS spectrum of a glycopeptide. However, compared to peptide sequences, glycosylation does not occur with linear addition of amino acid residues. Nevertheless, only a limited set of enzymes is involved in glycan biosynthesis for a given organism, and thus, using appropriate rules, a database of potential glycan structures can be limited to the biologically relevant forms.¹⁹ Moreover, as detailed below, since neither isomeric monosaccharides, nor linkages can be easily distinguished by low energy CID, the number of entries in the database is significantly lower than the number that are theoretically possible. Moreover, databases of known glycan structures⁴⁴ can be appended to the theoretical database.

Low energy CID fragmentation of glycans is not random, but predictable.³⁰ Several authors have investigated fragmentation of N-linked glycopeptides and concluded that (i) such glycopeptides are mainly fragmented along the glycosidic backbone;³⁰ while the monosaccharide rings as well as the peptide backbone remain intact;^{43, 45-49} (ii) the charge on the ion is mainly retained on the peptide moiety, and thus *Y-ion* fragments³⁰ are preferentially obtained; and (iii) fragment ions due to one or at most two cleavages per molecule are most often found.³⁰ Thus, given a glycan structure and a known peptide mass, a theoretical MS/MS spectrum can be generated. As a consequence, a specific database search algorithm can be developed to create theoretical glycopeptide spectra and calculate a similarity index to the experimental spectrum. As noted in the Introduction, while such software has been established, it is typically applicable only for annotation of individual MS/MS spectra and thus is not suitable for high throughput application. Here, we report on an alternative approach, where a peptide database search program, such as Sequest, is customized for selection of potential glycan structures. Importantly, the GlyDB approach can be broadly applied to other search programs as well that are commonly available.

In the GlyDB approach, the search is performed sequentially for all glycopeptides of interest. It is assumed that each glycopeptide can be potentially present in any glycoform in the database. Since the peptide mass and the list of all potential glycan structures are known, the masses of all glycoforms, within the restraints described above, can be readily calculated.

Analogous to peptide identification, for a given MS/MS spectrum, only a limited number of potential glycoforms will match the m/z value of the precursor ion within the specified range of mass tolerance. Moreover, due the small number of common monosaccharides with unique mass (5 compared to 19 amino acids), we estimated that there are roughly 1,000 non-isomeric

glycan compositions with molecular weight lower than 5,000 Da. Thus, compared to peptide database searching with thousands of unique peptide masses, most of the acquired MS/MS spectra will not match the mass of a given glycopeptide structure, resulting in a significant reduction in the number of tandem mass spectra that need to be compared with the database. This lower number of potential nonisomeric structures greatly reduces both the searching time and the likelihood of a random match. In a typical LC-MS analysis of a digest of a simple glycoprotein mixture, only a small fraction of acquired MS/MS spectra will thus match the mass of a glycopeptide structure, even in the case of broad mass tolerance, as required for ion trap instruments. Thus, for simple mixtures of glycoproteins, it is not necessary to pre-select MS/MS spectra of glycopeptides; instead all MS/MS spectra, including those for peptides acquired during the LC-MS run can be searched at the same time. For more complex mixtures, the glycopeptides can be recognized by their clustering at given elution time windows.

In order to utilize currently available peptide database search engines, the glycan structures are formatted into linear sequences codes consisting of single-letter amino acid abbreviations. Then, in order to generate masses of monosaccharides, permanent modifications of these selected amino acid residues are specified such that the modified masses correspond to the masses of these monosaccharides. For example, a single letter amino acid code H (histidine mass= 137) with a specified permanent modification (mass increase) of +25 will be interpreted as the symbol H with a mass of 162, corresponding to the mass of hexose. Table 1 lists (a) five major monosaccharides forming N-linked glycans and their symbols, (b) amino acid codes used to represent the monosaccharides, (c) molecular weights of monosaccharides and amino acid residues, and (d) masses of permanent modifications added to the original molecular weights of amino acid residues. Other monosaccharides can be included as needed by modifying the masses of additional amino acids. A special meaning is assigned to the symbol K (originally amino acid lysine). Since it is expected that under low energy CID conditions it is unlikely that the peptide moiety would fragment, the entire peptide sequence is represented by this single symbol. The permanent modification of K has must be separately specified for each individual peptide moiety of a glycopeptide to be searched.

It should be noted that since low energy CID fragmentation cannot distinguish isomeric monosaccharides, hexose (H) represent all isomeric hexoses such as mannose, galactose and glucose. Similarly, HexNAc (N) and deoxy-hexose (F) represent classes of other isomeric monosaccharides such as GlcNAc and fucose, respectively. For example, the mass of a hypothetical glycopeptide with a glycan moiety consisting of 4 linearly connected mannose units can be represented by the linear sequence HHHHK, where H is mannose and K represents the mass of the modified lysine corresponding to the mass of a peptide moiety. The molecular weight of this glycopeptide can be calculated by summing 4 times the mass of H and the mass of K minus. In addition, assuming fragmentation of a single glycosidic bond at a time, the same linear sequence HHHHK can be used by peptide database searching program to calculate the masses of *B*- and *Y*-ions, in a manner similar to that for peptide sequences.

Naturally occurring N-glycans are branched, not linear, and an experimental tandem mass spectrum of a glycopeptide will therefore contain peaks of fragments representing various branched structures. The fragment ions of a branched structure cannot be predicted directly from a *single linear* sequence. In order to overcome this limitation, we have developed a notation that allows for the generation of *multiple linear sequences* to represent fragmentation of branched glycans, thus allowing experimental spectra to be matched to a database of expected branched glycopeptide structures using a standard peptide database search program.

To illustrate the principles of branched annotation, consider a simple non-symmetrical biantennary glycopeptide. Figure 1A shows the structure of the glycopeptide as well as a schematic representation of conversion of the standard glycan notation to the GlyDB notation.

Figure 1B illustrates linearization of the branched structure. In order to facilitate visualization of linear structures, slashes have been added to represent separation of individual branches and the glycan core; however, these slashes are not used in the actual database. Since the glycopeptide is represented by linear sequences, its mass, as well as the masses of the ion fragments, can be automatically calculated by the database search program in a manner similar to that for linear glycans, i.e. summing of the masses of modified amino acids representing individual monosaccharides.

Low energy CID of glycopeptides leads preferentially to formation of *B*- and *Y*-ions, corresponding to cleavages at glycosidic bonds. Figure 1C visualizes the fragmentation process that will likely occur in the MS/MS spectrum of the glycopeptide in Fig 1A. Table 2 lists all potential *B*- and *Y*-ion fragments corresponding to cleavages of a single glycosidic bond. However, due to the higher ionization efficiency of fragments carrying the peptides moiety, *Y*-ions typically dominate the MS/MS spectrum. It can be seen from Fig. 1A that fragments at bonds 1 to 3 or 4 to 7, relate to CID cleavages that occur in one of the two branches, while fragments 8 and 9 correspond to cleavages in the core region of the glycan. Resulting linear sequences can then be used to calculate the mass of individual ion fragments. As an example, the linear sequence NH/SHNH/HNNK can be used to calculate the mass of *Y*-ion after bond 1 is cleaved. Similarly, if bond 2 is cleaved, the linear sequence H/SHNH/HNNK can represent the *Y*-ion formed. In general, the linear sequence HNH/SHNH/HNNK can be selected by the peptide database search program to generate theoretical fragment ions corresponding to cleavages at positions 1, 2 and 3. As can be further seen in Table 2, fragment ions due to cleavages at positions 8 and 9 will also be generated. In order to form fragments resulting from branch 2, i.e. fragments 4, 5 and 6, linear sequence, SHNH/HNH/HNNK, must also be in the database. In the case of more complex structures, such as tri- and tetra-antennary glycopeptides, additional linear sequences are added to the database to represent additional fragments resulting from the fragmentation in individual branches.

While the most common one bond cleavages have been considered in Table 2, cleavages of two glycosidic bonds per molecule are possible,³⁰ and the database search program must account for such fragments. Since *Y*-ions are preferentially observed, the cleavages that occur in a bond adjacent to the glycan core, i.e. bonds 3, 7, 8 and 9 in Fig. 1A, are equivalent to one bond cleavages. Thus, no new sequences have to be added to the database in this case. However, when two cleavages originate from different branches of the glycan, e.g. bonds 1 and 5, new fragments will be formed. It is clearly possible to generate linear sequences for all combinations fragmentation in different branches; however, this approach would lead to a large number of sequences with many fragments in common. Fortunately, it has been shown that fragments with one of the bond cleavages in the branch position adjacent to the core (e.g., bonds 2 or 6 in Fig. 1A) and the other cleavage in any bond of the remaining branches are most frequently observed.²⁰ As a result, only two additional sequences have been added to the database to account for cleavages of two bonds, see Supplemental Fig S1 and S2.

Table 3, summarizes all linear sequences that are available in GlyDB for the glycopeptides discussed above. Sequence HNSHNHHHNNK corresponds to one cleavage at bond 2 (bond adjacent to the core), with the second cleavage occurring elsewhere in the glycan. Similarly, sequence SHNHNHHHNNK can produce ions corresponding to two cleavages per molecule, i.e. cleavage at bond 6 and at another bond position in the glycan.

Additional linear sequences have been included in the database to account for glycan structure features such as bisecting GlcNAc, as shown in Table 3. Other common patterns of two bond cleavages occur in the case of glycans with a fucose at the reducing terminal. Loss or rearrangement of the fucose is commonly observed;⁵⁰ thus, sequences representing these structures (loss and rearrangement) have been added to the database, as shown in Table 3 (see

also Supplemental Fig S1). It should be noted that fragment ions corresponding to three or more bond cleavages are not considered, as they are not likely to occur in low energy CID. Furthermore, it should be emphasized that the database search program will generate m/z values for all fragments in all appropriate charge states within the mass range. Currently, the glycan database contains about 5,000 glycan structures represented by approximately 14,000 linear sequences. The glycan structures are derived based on the most common N-linked glycans found, see Supplemental Material.⁵¹ In the future, the database can clearly be expanded, including glycans from experimental databases and other species, e.g. plants.

In order to perform a database search, linear sequences representing a given glycoform have been combined in a single entry in a fasta format⁵², analogous to protein sequences stored in protein databases, see Fig. 1D. The fasta format consists of a header line that starts with a greater than sign ">" and protein name followed by line(s) with protein sequence represented with single letter amino acid abbreviations. In the case of GlyDB, the header line consists of the glycan identifier followed by description of the glycan. The next line is a single string consisting of concatenated multiple linear sequences.

Since the peptide moiety is represented in individual linear sequences by lysine (K), using Lys-C enzymatic specificity in the database search results in splitting of the concatenated sequence into individual linear sequences that are then automatically matched to the experimental spectra. Thus, the linear sequences can be viewed as peptides that all belong to the same glycoform. The fasta format of the GlyDB database allows any protein database search engine to be used for glycoform assignment, in this work Sequest.

Compared to peptides, glycopeptides do not show a significant dependence of fragmentation on the charge state of the precursor ion nor on the length of the peptide moiety, as the number of theoretical glycan fragments will depend solely on the complexity of the glycan structure. In addition, since the MS/MS spectra of glycopeptides will typically be dominated by *Y-ion* fragments, matching scores for glycopeptides will generally be lower than scores commonly encountered in peptide database searching, where typically both *b-* and *y-ion* series occur. As a result, criteria applied for selection of significant matches have to be defined. Based on preliminary studies with known structures, it has been determined that, as a rule, linear glycan sequences with an Xcorr value greater than 1.5 should be selected for further examination. The MS/MS spectra of such potential glycoforms are then manually inspected for consecutive *Y-ion* series distinctive from noise. In addition, chromatographic retention time is taken into account since glycopeptides, that are above roughly 5 amino acids, derived from the same glycosylation site generally elute close to one another in reversed phase chromatography,⁴⁰ and thus multiple glycopeptides from the same backbone will often appear within a several minute elution window.

As shown in the Results and Discussion Section, the assignment of the most likely isomeric structure to a particular MS/MS spectrum relies on the power of multiple linear sequence matches. Because a single glycan structure will be represented by such multiple linear sequences, in the case of a correct match, more than one linear sequence will be assigned to the same MS/MS spectrum. Although some ions of other isomeric glycans may match fragment ion peaks of the true glycan, the degree of matching, and thus the Xcorr value should be lower for incorrect matches. If, however, multiple structures can match the same spectrum with similar and significant Xcorr values, all structures will be reported. In all cases, manual inspection must be employed to determine the most likely match or alternatively, selective fragment ions must be targeted in an additional experiment using MSⁿ to reveal the identity of the ions.

Experimental

Reagents

Sequencing grade trypsin was purchased from Promega (Madison, WI). Bovine fetuin, lectin from *Erythrina cristagalli*, human transferrin, human lactoferrin, human immunoglobulin G (IgG) and epidermal growth factor receptor (EGFR), as well as dithiothreitol (DTT), iodoacetamide (IAA), guanidine hydrochloride, and ammonium bicarbonate, were obtained from Sigma (St. Louis, MO). Formic acid, acetone and acetonitrile were purchased from Fisher Scientific (Fair Lawn, NJ), and HPLC grade water was from J. T. Baker (Bedford, MA). The oligosaccharide, whose structure is shown in Fig. 3A, was purchased from Prozyme (San Leandro, CA, USA). Two aminooxyacetyl modified peptides NH₂-O-CH₂CO-HNETWDSLQR (purity 80%) and NH₂-O-CH₂CO-HNETWDSLQSHNETR (purity 90%) were custom synthesized by Sigma-Genosys (Woodlands, TX, USA).

Synthesis of glycopeptides

Two glycopeptides were synthesized in-house, according to,²⁷ using commercial oligosaccharides and the custom synthesized peptides. Briefly, 2 µL (200 pmol) of an oligosaccharide was added to 0.5 µL (200 pmol) of an aqueous solution of an aminooxyacetyl peptide, mixed with 30 µL of CH₃CN and incubated overnight at 37 °C. The reaction product was vacuum dried and used without further purification. While the yield of the glycopeptide was not high (10%), the glycan structure determination could nevertheless be readily determined.

MS analysis of synthetic glycopeptides

The synthetic glycopeptides were dissolved in 0.1% formic acid in 50% methanol to a concentration of 5 µM and directly infused into a linear ion trap LTQ MS (Thermo Fisher, San Jose, CA, USA). All MS and MS/MS spectra were acquired using static nanospray in the manual mode.

Enzymatic digestion of glycoproteins

1 mg of a glycoprotein was dissolved in 6 M guanidine hydrochloride and 100 mM NH₄HCO₃. The glycoprotein was reduced with 20 mM DTT for 30 min at 37 °C and alkylated with 50 mM of IAA in the dark for 1.5 h at room temperature. After desalting using a Microcon spin column (10 kDa MWCO; Millipore, Bedford MA), trypsin (1:25 w/w) was added to the sample and incubated at 37 °C overnight. Digestion was stopped by addition of 1% formic acid. A C₁₈ ZipTip (Millipore, Bedford, MA, USA) was used for sample cleanup. *E. cristagalli* lectin was digested using the same protocol but without reduction and alkylation.

LC-MS

LC-MS experiments of glycoprotein digests were performed on an LTQ MS or LTQ FT MS instrument (Thermo Fisher) with an Ultimate nanoLC system (Dionex, Mountain View, CA), using a reversed phase column (75 µm i.d. × 10 cm, BioBasic C₁₈, 5 µm particle size, Thermo Fisher) at a flow rate of 200 nL/min and 30 or 40 min linear gradient. Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1% formic acid in acetonitrile. The mass spectrometer was operated in the data-dependent mode using one full MS scan, followed by sequential MS/MS scans for the four most intense ions. Dynamic exclusion was employed with duration of 60 s.

Epidermal growth factor receptor dataset (EGFR)

Previously generated data files were used.⁵³ Briefly, reduced and alkylated EGFR (1 pmol) was digested using Lys-C and analyzed by LC-MS with the hybrid LTQ-FT MS instrument.

Database search and glycan structure assignment

N-glycan structures were identified by automatic searching of the raw data against the custom-built glycan structure database (GlyDB), using the Sequest algorithm (BioworksBrowser, version 3.2, Thermo Fisher). The Sequest search was conducted with a mass tolerance of 1.4 Da for the precursor mass, and 1 Da for the fragment ions.

Results and Discussion

In order to evaluate the performance of GlyDB strategy, glycopeptides from well characterized glycoproteins and a synthetic glycopeptide with a known glycan structure were examined. The ability of the GlyDB approach to detect glycopeptides in a proteolytic digest of a glycoprotein was initially examined, and then the capability of the approach to differentiate several isomeric glycopeptides with the same glycan composition was explored. Finally, comprehensive characterization of two glycosylation sites in bovine fetuin as well as the epidermal growth factor receptor was performed.

Automated Selection of Glycopeptide MS/MS Spectra

In a typical LC-MS analysis of a glycoprotein digest using data dependent MS/MS acquisition, spectra of peptides, glycopeptides as well as contaminants and noise are collected, and thus selection of those spectra that correspond to glycopeptides can be a challenging, time consuming task. Since the GlyDB approach utilizes the mass of the peptide moiety and a database of potential glycan structures, for a simple mixture of digested glycoproteins it is not necessary to preselect MS/MS spectra of glycopeptides prior to the search because only those spectra with a precursor ion m/z value within the mass tolerance of a potential glycopeptide will be selected for matching to the theoretically predicted fragmentation pattern. In order to demonstrate this ability of GlyDB to select automatically glycopeptide MS/MS spectra, we processed LC-MS data from the analysis of a tryptic digest of a known glycoprotein, and then repeated the search using several incorrect masses of one of the peptide moieties to check for false positives.

A tryptic digest of a lectin isolated from *E. cristagalli*, which is known to have only one major glycoform,⁴⁰ was analyzed by LC-LTQ MS using a 30 minute gradient in which a total of 4,324 MS/MS spectra were collected. All MS/MS spectra were searched against the GlyDB glycoform library using the Sequest search program. Only one hit with an Xcorr score greater than 1.5 was found, and this hit (Xcorr = 2.2) corresponded to the known glycan structure. Figure 2 shows assignment of fragment ions. As expected, the majority of fragments correspond to *Y-ions* in 2+ and 3+ charge states. Importantly, the database search program, Sequest in this case, generates fragments with all potential charge states, which not only improves the scoring but also makes the annotation easier. This is also a significant advantage of the GlyDB approach compared to similar programs for analysis of MS/MS spectra because these programs usually consider only single or double charge ions. Thus, deconvolution of the mass spectrum would be necessary, which is not possible for low resolution mass spectra acquired using an ion trap mass spectrometer.

The above result illustrates that even though all collected MS/MS spectra were analyzed by the database search program, only the spectrum corresponding to the correct glycoform generated a score above the significant threshold. The success of this approach is due to the specificity of the GlyDB approach to select only those MS/MS spectra with a precursor ion within the mass tolerance of potential glycopeptides and the matching of MS/MS spectra to predicted fragments corresponding to this glycoform. In order to confirm that the above assignment was not random, we carried out 8 additional searches using the same database search parameters but with incorrectly specified masses of the peptide moiety. Such

glycopeptides do not exist in the sample and thus no significant match should be observed. We used a peptide mass increment of 5 Da, up to a molecular weight 20 Da lower and higher than real mass of the peptide, i.e. -20, -15, -10, -5, +5, +10, +15 and +20 Da. All of these searches yielded assignments with Xcorr scores below 1.5 (data not shown). From this result, it can be concluded that for simple glycopeptide mixtures, when the number of collected MS/MS spectra is on the order of a few thousand, there is no need to focus on spectra corresponding to glycopeptides; all spectra can be interrogated. On the other hand, for more complex mixtures of digested proteins with hundreds of thousands of MS/MS spectra, higher rates of false positive hits could occur. In such cases, additional information has to be incorporated to narrow down the potential elution window for a particular glycopeptide. For example, due to the close elution of glycopeptides and their deglycosylated forms in the reverse phase LC, the same mixture can be analyzed after deglycosylation to determine the specific elution time windows. Alternatively, new MS instrumentation such as electron transfer dissociation (ETD) can be selected to determine the elution position of the glycopeptide(s).⁴¹

Identification of glycan structure of synthetic glycopeptide

The ability of GlyDB approach to differentiate isomeric glycan structures was next examined. In order to evaluate this feature, we synthesized a glycopeptide using a purified glycan and an activated peptide, aminooxyacetyl modified peptides: $\text{NH}_2\text{-O-CH}_2\text{CO-HNETWDSLQR}$. The experimental MS/MS spectrum obtained by direct infusion of this synthetic glycopeptide into the linear ion trap mass spectrometer was searched against a custom GlyDB database consisting of five isomeric glycan structures with the same glycan composition as the synthetic glycopeptide, see Fig. 3A. It should be noted that some of the structures in this figure were designed for illustration and may not naturally occur.

Results of the Sequest search of the synthetic glycopeptide are presented in Table 4. The value of Xcorr can be seen to be greater than 1.5 for at least one linear sequence for all five potential glycopeptide structures, indicating the presence of glycosidic fragments common for all structures. The GlyDB approach thus allowed selection of potential matching structures, but further manual annotation and validation was clearly necessary. To illustrate this further, Figure 3B shows the annotation of the experimental spectrum using the structure with the highest Xcorr value, i.e. structure 1. Note that after manual annotation, nearly all of the 23 most intense peaks in the MS/MS spectrum were assigned to *Y*-ions of 2+ or 3+ suggesting a potential match.

It is necessary to compare examine ion assignments for individual glycoforms in order to confirm the match. For example, the peak at m/z 1138.8 was assigned to the linear sequence, FNK, indicating the fucosylation was at the reducing terminus. Interestingly, both of the structures with core fucosylation (structures 1 and 3 in Fig. 3A) had higher Xcorr scores compared to glycopeptides where the fucose was attached to one of the branches, see Table 4 and Fig. 3B. On the other hand, using annotation for structure 2 (data not shown), the peak at m/z 1138.8 did not match any predicted ion, further confirming that the spectrum likely corresponded to a glycan with a core fucose. On this basis, structures 2, 4 and 5 were excluded from further consideration. The remaining two structures, i.e. 1 and 3, differ by the position of a HexNAc residue. Peaks at m/z 948.8 and 1422.8, can be attributed to a linear sequence NHHNHNFNK, present only in the case of a bisecting HexNAc. This result indicates that the glycan has a bisecting HexNAc and that structure 1 is again the most likely candidate. Note that while the structure with the highest Xcorr score was assigned to linear sequences corresponding to the correct glycopeptide, this need not always be the case.

It should be noted that the individual linear sequences could also be viewed as preferred low energy CID fragmentation pathways for a particular glycan structure. More likely fragmentation pathways will produce more intense fragment ions leading to higher Xcorr values. For example, linear sequences **a**) or **b**) in Fig 3B can be used to produce a set of

fragments that can account for most of the ion current, thus likely representing the most favorable fragmentation pathway and also leading to the highest Xcorr score. Additionally, structures **c**) and **d**) help to explain less likely fragments resulting from less favorable fragmentation pathways, and thus contributing to selection of the most appropriate isomeric structure.

Interestingly, a standard Sequest search for the synthetic glycopeptide showed the presence of the same peptide sequence in more than one protein, indicating that the given peptide was not unique. For example, referring to Table 4 in the column *other matches*, the linear sequence FNNHHNHNHNK, associated with glycan number 1, is present in two other glycoforms in the database, which means that this linear sequence cannot be used to differentiate between these structures. Nevertheless, this example illustrates that a GlyDB search can be used to assign different Sequest scores to isomeric glycan structures, and that these scores can be used in the annotation of the glycan structures. However, manual validation is necessary, and GlyDB serves the role to minimize the number of candidates necessary for further examination, thus reducing the time necessary to interpret the MS/MS spectra.

Analysis of bovine fetuin – multiple glycosylation sites

Bovine fetuin is a well-characterized glycoprotein with 3 N-linked and 4 O-linked glycosylation sites.⁵⁴ In order to evaluate the performance of GlyDB for the identification of multiple glycosylation sites, a tryptic digest of bovine fetuin was analyzed by LC-MS using the LTQ MS. Two N-linked glycosylation sites at positions N-156 and N-176 are selected to illustrate the analysis of a protein with multiple glycosylation sites. In order to determine the MS/MS spectrum corresponding to the **N-156** glycosylation site, the GlyDB search was conducted using the addition of 1594.8 Da to K to represent the fetuin tryptic glycopeptide LCPDCPLLAPLNSR. All the MS/MS spectra were then searched against the full GlyDB database, and the results of the search showed that significant hits (Xcorr > 1.5) were in the chromatographic time window of 30–34 minutes, indicating the presence of a closely eluting family of glycopeptides. In total, 6 structures, see Fig. 4A, with the highest Xcorr scores for a given MS/MS spectrum were assigned. See Fig. 4B for an example of an annotated MS/MS spectrum of triantennary glycan with three sialic acids. Similar to Fig. 3B, fragments with multiple charge ions, 2+, 3+ and 4+ appear in the spectrum. All 6 structures were found to be consistent with fetuin glycosylation.⁵⁴ Unfortunately, the exact location of the sialic acid residues at individual branches could not be determined from the low energy CID spectra, and thus ambiguous positions are indicated in Fig 4A.

Next, the **N-176** glycosylation site was next examined in a similar manner by searching with a K modification of 2870.4 Da, corresponding to the peptide VVHAVEVALATFNAESNGSYLQLVEISR. In total, three major glycosylation structures were automatically assigned to this glycosylation site (Xcorr > 1.5). All three glycoforms eluted in the 36–39 min elution window, and these matches were again consistent with reported structures.⁵⁴ Interestingly, several MS/MS spectra were associated with two different compositions with either two sialic acid or one sialic acid and two fucoses in terminal branches (data not shown). Both structures had highly significant hits (Xcorr for best match 3.6 and 3.4, respectively). Using manual inspection, it was confirmed that the former two sialic side glycoform was the correct identification. The exact molecular mass of glycopeptide, as calculated from a high resolution MS spectrum acquired using an LTQ-FT MS instrument, corresponded to a triantennary structure with two sialic acids rather than one sialic acid and two fucoses (difference of app. 1 Da). Importantly, in such cases where multiple significant hits are found, GlyDB will report all structures above an Xcorr score threshold. The user must then manually determine the most likely glycoform. This example again illustrates that GlyDB is very useful for selection and annotation of MS/MS spectra; however, manual evaluation of

the results is still often necessary. Nevertheless, the GlyDB approach significantly simplifies the process of selecting high quality MS/MS spectra corresponding to glycopeptides and provides guidance in assigning putative glycan structures to these spectra.

Analysis of epidermal growth factor receptor (EGFR)

As a final example, we examined LC-MS data of Lys-C digest of the epidermal growth factor receptor (EGFR) that was originally manually annotated.⁵³ EGFR is a glycoprotein with 12 theoretical N-glycosylation sites, and we selected two sites,⁵³ i.e. N-337 and N-420, to illustrate the GlyDB search. Even though the GlyDB search could be conducted against all collected MS/MS spectra, we limited the search to the previously determined retention time windows to reduce the number spectra that must be examined manually.⁵³ Since the majority of incorrectly assigned spectra correspond to spectra with a high level of noise, removal of low quality MS/MS spectra prior to the database searching would allow for analysis using all collected MS/MS spectra without the need for extensive manual validation of matches. In this work, GlyDB analysis was performed in 6 minute retention time windows: 37–42 mins for N-337 and 40–45 minutes for N-420. The partial overlap of these two retention windows had no effect on the GlyDB search, since different masses of the peptide moieties were used.

The database search results for site N-337 are presented in Table 5A. In agreement with previous work⁵³, the most likely matches corresponded to high hexose (mannose) type glycans. Specifically, structures 1 to 4 represent a family of Man5 to Man8 glycopeptides. High Xcorr values, as well as, the matching of the 4+ and 5+ charge states, determined from the LTQ-FT MS spectra, to identical glycans strongly suggested a correct assignment.

The additional match of glycan structures with Xcorr greater than 1.5 originally assigned to a complex glycan (match 5 in Table 5A) was manually examined. It was found that this match was likely random since *Y-ions* were matched to peaks close to the noise level. In the case of the N-420 glycosylation site, Table 5B, the most likely match, a bisialo biantennary glycopeptide, corresponded to the same structure obtained by manual annotation.⁵³ Matches 2 and 3 in Table 5B were excluded, based on the rather low quality of the MS/MS spectra. The same MS/MS spectrum was assigned to matches 1 and 4. As shown above, these two glycopeptides contain either two sialic acids (match 1) or one sialic acid and two fucoses (match 4), representing a mass difference of roughly 1 Da. Match 1 had a significantly higher value of Xcorr; in addition, only one linear sequence agreed with for glycopeptide 5 compared to 4 sequences for structure 1. It was therefore concluded that glycopeptide 1 was the likely assignment. Additionally, accurate mass of the precursor ion could be used to eliminate match 5. In summary, these examples for EGFR illustrate that GlyDB is a useful tool that allows selection of potential glycopeptide MS/MS spectra and assignment to potential glycoforms. Importantly, using the GlyDB approach, the selection of MS/MS spectra and assignment of individual fragments with multiple charge states is done automatically saving hours of time that would be otherwise requires for manual annotation. In addition, multiple glycosylation sites can be sequentially evaluated, and glycosylation at a specific site can be determined.

Conclusions

We have developed a new procedure for annotation of low resolution MS/MS spectra of glycopeptides arising from low energy CID spectra. The procedure removes the burden of time consuming manual search for MS/MS spectra of glycopeptide and their interpretation. The approach is based on matching, using the search engine Sequest, of experimental spectra to theoretical glycopeptide spectra generated by a linearized database of glycan structures. For relatively pure glycoproteins, we have demonstrated that GlyDB can reliably find glycopeptides from a few thousand collected MS/MS spectra and assign the correct glycan structures. Furthermore, these structures are linked to specific glycosylation sites, which could

be lost using a deglycosylation strategy with subsequent characterization of released glycans. The GlyDB approach will be very beneficial for evaluation of spectra acquired by LC-MS using the combination of CID and ETD fragmentation. The CID MS/MS spectra can lead to assignment of the specific glycoform, while the ETD spectra can be utilized to confirm the sequence of the peptide moiety and the site of glycosylation.

The GlyDB approach was primarily developed for annotation of low resolution spectra generated by ion trap mass spectra; however, the approach would be a powerful tool with high mass accuracy in both the MS and MS/MS modes. For example, a high resolution MS scan could be used to determine accurately a specific glycan composition, thus minimizing false positive matches. The GlyDB approach removes the burden of time consuming manual annotation of glycopeptide spectra. As demonstrated in this work, in the case of simple structures, such as biantennary glycans, the GlyDB approach can provide direct assignment. In the case of more complex glycan structures, GlyDB typically selects several potential candidate structures, but in order to assign the most likely structure, additional experiments may be required. Nevertheless, a list of likely candidates will reduce the time necessary for determining the correct glycoform.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

This work was supported by National Institutes of Health grant GM 15847. The authors appreciate the help of Drs. Y. Wang and J. Kim, and Ms. L. Cao for some of the LC-MS data files used for testing. Contribution number 888 from the Barnett Institute.

References

1. Raman R, Raguram S, Venkataraman G, Paulson JC, Sasisekharan R. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat Methods* 2005;2(11):817–24. [PubMed: 16278650]
2. Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999;1473(1):4–8. [PubMed: 10580125]
3. Woods RJ, Edge CJ, Dwek RA. Protein surface oligosaccharides and protein function. *Nat Struct Biol* 1994;1(8):499–501. [PubMed: 7664073]
4. Varki A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* 1993;3(2):97–130. [PubMed: 8490246]
5. Hakomori S. Glycosylation defining cancer malignancy: new wine in an old bottle. *Proc Natl Acad Sci U S A* 2002;99(16):10231–3. [PubMed: 12149519]
6. Rudd PM, Elliott T, Cresswell P, Wilson IA, Dwek RA. Glycosylation and the immune system. *Science* 2001;291(5512):2370–6. [PubMed: 11269318]
7. Beckers J, Herrmann F, Rieger S, Drobyshev AL, Horsch M, Hrabe de Angelis M, Seliger B. Identification and validation of novel ERBB2 (HER2, NEU) targets including genes involved in angiogenesis. *Int J Cancer* 2005;114(4):590–7. [PubMed: 15609325]
8. Fuessel S, Sickert D, Meyer A, Klenk U, Schmidt U, Schmitz M, Rost AK, Weigle B, Kiessling A, Wirth MP. Multiple tumor marker analyses (PSA, hK2, PSCA, trp-p8) in primary prostate cancers using quantitative RT-PCR. *Int J Oncol* 2003;23(1):221–8. [PubMed: 12792797]
9. Peracaula R, Tabares G, Royle L, Harvey DJ, Dwek RA, Rudd PM, de Llorens R. Altered glycosylation pattern allows the distinction between prostate-specific antigen (PSA) from normal and tumor origins. *Glycobiology* 2003;13(6):457–70. [PubMed: 12626390]

10. Norum LF, Erikstein B, Nustad K. Elevated CA125 in breast cancer--A sign of advanced disease. *Tumour Biol* 2001;22(4):223–8. [PubMed: 11399947]
11. Kirmiz C, Li B, Joo An H, Clowers BH, Chew HK, Lam KS, Ferrige A, Alecio R, Borowsky AD, Sulaimon S, Lebrilla CB, Miyamoto S. A serum glycomics approach to breast cancer biomarkers. *Mol Cell Proteomics*. 2006
12. Varki, A. *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY: 1999. p. xviip. 653
13. Fuster MM, Esko JD. The sweet and sour of cancer: glycans as novel therapeutic targets. *Nat Rev Cancer* 2005;5(7):526–42. [PubMed: 16069816]
14. An HJ, Miyamoto S, Lancaster KS, Kirmiz C, Li B, Lam KS, Leiserowitz GS, Lebrilla CB. Profiling of glycans in serum for the discovery of potential biomarkers for ovarian cancer. *J Proteome Res* 2006;5(7):1626–35. [PubMed: 16823970]
15. Lamari FN, Kuhn R, Karamanos NK. Derivatization of carbohydrates for chromatographic, electrophoretic and mass spectrometric structure analysis. *J Chromatogr B Analyt Technol Biomed Life Sci* 2003;793(1):15–36.
16. Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res* 2003;31(1):511–3. [PubMed: 12520065]2003 update
17. Zhen Y, Caprioli RM, Staros JV. Characterization of glycosylation sites of the epidermal growth factor receptor. *Biochemistry* 2003;42(18):5478–92. [PubMed: 12731890]
18. Clerens S, Van den Ende W, Verhaert P, Geenen L, Arckens L. Sweet Substitute: a software tool for in silico fragmentation of peptide-linked N-glycans. *Proteomics* 2004;4(3):629–32. [PubMed: 14997486]
19. Goldberg D, Sutton-Smith M, Paulson J, Dell A. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* 2005;5(4):865–75. [PubMed: 15693066]
20. Harvey DJ. Electrospray mass spectrometry and fragmentation of N-linked carbohydrates derivatized at the reducing terminus. *J Am Soc Mass Spectrom* 2000;11(10):900–15. [PubMed: 11014452]
21. Lapadula AJ, Hatcher PJ, Hanneman AJ, Ashline DJ, Zhang H, Reinhold VN. Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MS(n) data. *Anal Chem* 2005;77(19):6271–9. [PubMed: 16194088]
22. Ethier M, Krokhin O, Ens W, Standing KG, Wilkins JA, Perreault H. Global and site-specific detection of human integrin alpha 5 beta 1 glycosylation using tandem mass spectrometry and the StrOligo algorithm. *Rapid Commun Mass Spectrom* 2005;19(5):721–7. [PubMed: 15702487]
23. Mizuno Y, Sasagawa T, Dohmae N, Takio K. An automated interpretation of MALDI/TOF postsource decay spectra of oligosaccharides. 1. Automated peak assignment. *Anal Chem* 1999;71(20):4764–71. [PubMed: 10546541]
24. Morelle W, Michalski JC. Sequencing of oligosaccharides derivatized with benzylamine using electrospray ionization-quadrupole time of flight-tandem mass spectrometry. *Electrophoresis* 2004;25(14):2144–55. [PubMed: 15273998]
25. Wuhler M, Deelder AM, Hokke CH. Protein glycosylation analysis by liquid chromatography-mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci* 2005;825(2):124–33.
26. 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]
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. Mechref Y, Novotny MV, Krishnan C. Structural characterization of oligosaccharides using MALDI-TOF/TOF tandem mass spectrometry. *Anal Chem* 2003;75(18):4895–903. [PubMed: 14674469]
29. Mechref Y, Kang P, Novotny MV. Differentiating structural isomers of sialylated glycans by matrix-assisted laser desorption/ionization time-of-flight/time-of-flight tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2006;20(8):1381–9. [PubMed: 16557638]
30. Domon B, Costello CE. A Systematic Nomenclature for Carbohydrate Fragmentations in FAB/MS of Glycoconjugates. *Glycoconjugate J* 1988;5:397–409.

31. Harvey DJ. Fragmentation of negative ions from carbohydrates: part 1. Use of nitrate and other anionic adducts for the production of negative ion electrospray spectra from N-linked carbohydrates. *J Am Soc Mass Spectrom* 2005;16(5):622–30. [PubMed: 15862764]
32. Harvey DJ. Fragmentation of negative ions from carbohydrates: part 2. Fragmentation of high-mannose N-linked glycans. *J Am Soc Mass Spectrom* 2005;16(5):631–46. [PubMed: 15862765]
33. Harvey DJ. Fragmentation of negative ions from carbohydrates: part 3. Fragmentation of hybrid and complex N-linked glycans. *J Am Soc Mass Spectrom* 2005;16(5):647–59. [PubMed: 15862766]
34. Zhang H, Singh S, Reinhold VN. Congruent strategies for carbohydrate sequencing. 2. FragLib: an MS(n) spectral library. *Anal Chem* 2005;77(19):6263–70. [PubMed: 16194087]
35. Goldberg D, Bern M, Li B, Lebrilla CB. Automatic determination of O-glycan structure from fragmentation spectra. *J Proteome Res* 2006;5(6):1429–34. [PubMed: 16739994]
36. An HJ, Tillinghast JS, Woodruff DL, Rocke DM, Lebrilla CB. A new computer program (GlycoX) to determine simultaneously the glycosylation sites and oligosaccharide heterogeneity of glycoproteins. *J Proteome Res* 2006;5(10):2800–8. [PubMed: 17022651]
37. Gaucher SP, Morrow J, Leary JA. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal Chem* 2000;72(11):2331–6. [PubMed: 10857602]
38. Ashline D, Singh S, Hanneman A, Reinhold V. Congruent strategies for carbohydrate sequencing. 1. Mining structural details by MS(n). *Anal Chem* 2005;77(19):6250–62. [PubMed: 16194086]
39. Block TM, Comunale MA, Lowman M, Steel LF, Romano PR, Fimmel C, Tennant BC, London WT, Evans AA, Blumberg BS, Dwek RA, Mattu TS, Mehta AS. Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proc Natl Acad Sci U S A* 2005;102(3):779–84. [PubMed: 15642945]
40. Hakansson K, Cooper HJ, Emmett MR, Costello CE, Marshall AG, Nilsson CL. Electron capture dissociation and infrared multiphoton dissociation MS/MS of an N-glycosylated tryptic peptide to yield complementary sequence information. *Anal Chem* 2001;73(18):4530–6. [PubMed: 11575803]
41. Lubeck, M.; Hartmer, R.; Koeleman, CAM.; Stam, J.; Brekenfeld, A.; Catalina, MI.; Deelder, AM.; Baessmann, C.; Ingendoh, A.; Wuhler, M. Improved Characterization of Glycopeptides using CID and ETD Ion Trap MS/MS. Third International Symposium on Separation and Characterization of Natural and Synthetic Macromolecules, Amsterdam, 2007; Amsterdam. 2007;
42. An HJ, Peavy TR, Hedrick JL, Lebrilla CB. Determination of N-glycosylation sites and site heterogeneity in glycoproteins. *Anal Chem* 2003;75(20):5628–37. [PubMed: 14710847]
43. Dell, A.; Morris, HR. Glycoprotein structure determination by mass spectrometry.; *Science*. 2001. p. 2351–6.
44. Consortium for functional glycomics In www.functionalglycomics.org
45. Hogan JM, Pitteri SJ, Chrisman PA, McLuckey SA. Complementary structural information from a tryptic N-linked glycopeptide via electron transfer ion/ion reactions and collision-induced dissociation. *J Proteome Res* 2005;4(2):628–32. [PubMed: 15822944]
46. Demelbauer UM, Zehl M, Plematl A, Allmaier G, Rizzi A. Determination of glycopeptide structures by multistage mass spectrometry with low-energy collision-induced dissociation: comparison of electrospray ionization quadrupole ion trap and matrix-assisted laser desorption/ionization quadrupole ion trap reflectron time-of-flight approaches. *Rapid Commun Mass Spectrom* 2004;18(14):1575–82. [PubMed: 15282782]
47. Janet Irungu DSD, Eden P. Go, Hui Jiang Hy-Vy Ha, George R. Bousfield, Desaire H. Method for Characterizing Sulfated Glycoproteins in a Glycosylation Site-Specific Fashion, Using Ion Pairing and Tandem Mass Spectrometry. *Anal Chem*. 2006
48. Peterman SM, Mulholland JJ. A Novel Approach for Identification and Characterization of Glycoproteins Using a Hybrid Linear Ion Trap/FT-ICR Mass Spectrometer. *J Am Soc Mass Spectrom* 2006;17(2):168–79. [PubMed: 16406561]
49. Itoh S, Kawasaki N, Harazono A, Hashii N, Matsuishi Y, Kawanishi T, Hayakawa T. Characterization of a gel-separated unknown glycoprotein by liquid chromatography/multistage tandem mass spectrometry: analysis of rat brain Thy-1 separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis. *J Chromatogr A* 2005;1094(12):105–17. [PubMed: 16257296]

50. Harvey DJ, Mattu TS, Wormald MR, Royle L, Dwek RA, Rudd PM. "Internal residue loss": rearrangements occurring during the fragmentation of carbohydrates derivatized at the reducing terminus. *Anal Chem* 2002;74(4):734–40. [PubMed: 11866052]
51. Kornfeld R, Kornfeld S. Assembly of asparagine-linked oligosaccharides. *Annu Rev Biochem* 1985;54:631–64. [PubMed: 3896128]
52. http://en.wikipedia.org/wiki/Fasta_format In
53. Wu SL, Kim J, Hancock WS, Karger B. Extended Range Proteomic Analysis (ERPA): a new and sensitive LC-MS platform for high sequence coverage of complex proteins with extensive post-translational modifications-comprehensive analysis of beta-casein and epidermal growth factor receptor (EGFR). *J Proteome Res* 2005;4(4):1155–70. [PubMed: 16083266]
54. Green ED, Adelt G, Baenziger JU, Wilson S, Van Halbeek H. The asparagine-linked oligosaccharides on bovine fetuin. Structural analysis of N-glycanase-released oligosaccharides by 500-megahertz ¹H NMR spectroscopy. *J Biol Chem* 1988;263(34):18253–68. [PubMed: 2461366]

Figure 1.

GlycDB linear representation of a branched glycan with a single sialic acid. (A) Conversion of standard glycan symbols to letters. Numbering of glycosidic bonds corresponds to the fragments listed in Table 2. (B) Examples of two potential linearizations of the structure in (A), dashed lines illustrate the process of linearization. Three potential fragmentation schemes and (C) the corresponding B- and Y- ions represented in linearized form. (D) Representation of glycopeptide in a fasta format suitable for database searching using conventional peptide database search programs. All four linear sequences representing the same glycan structure were concatenated into a single sequence.

Symbols: □ N – HexNAc (GlcNAc, GalNAc), ● H – Hex (Gal, Glc, Man), ▲ F – deoxyHex (fuc), ◇ S – Sialic acid (Neu5Ac), and K – peptide. Note: Slashes in the linear sequences were added to visualize individual branches; however, these symbols are not used in the actual database.

Figure 2.
Analysis of lectin from *E. cristagalli*. Automatically annotated MS/MS spectrum of a glycopeptide with a known glycan structure.
★ - xylose, other symbols as in Fig. 1.

N-176 site

N-156 site

Figure 4.

Analysis of bovine fetuin. **(A)** Glycan structures identified in LC-MS analysis of a fetuin tryptic digest. The star indicates structures reported in ref⁵⁴. **(B)** Example of an MS/MS spectrum of one glycoform of bovine fetuin automatically identified by GlyDB approach with annotation of individual fragments.

Symbols as in Fig. 1.

Table 1
Representation of monosaccharides and their masses in the GlyDB database

Monosaccharide	Symbol	Letter	Mass of monosaccharide (A)	Mass of amino acid (B)	Mass of modification (A-B)
Deoxyhexose (fucose)		F	146	147	-1
Hexose (mannose, glucose, galactose)		H	162	137	+25
HexNAc (GlcNAc, GalNAc)		N	203	114	+89
Sialic acid (Neu5Ac)		S	291	87	+204
Pentose (xylose)		Y	132	163	-31
Peptide	Pep	K	Peptide mass	128	Peptide mass -128 - 18*

* Mass of modification added to the Lys residue corresponds to the mass of peptide moiety minus the mass of lysine (128 Da) minus the mass of water (18 Da).

Table 2

Potential fragment ions and corresponding linear sequences for the glycan structure in Figure 1A, assuming only one CID cleaved bond.

Cleaved bond	Structure of <i>Y-ion</i> ^a	Linear form of <i>B-ion</i>	Linear form of <i>Y-ion</i>
1		H	NH/SHNH/HNNK
2		HN	H/SHNH/HNNK
3		HNH	/SHNH/HNNK
4		S	HNH/HNH/HNNK
5		SH	NH/HNH/HNNK
6		SHN	H/HNH/HNNK
7		SHNH	/HNH/HNNK
8		SHNH/HNH/H	NNK
9		SHNH/HNH/HN	NK

^aOnly structures of *Y-ions* are shown for clarity. Ion charges are omitted. Slashes visually indicate branches and core of the glycan. Symbols (H, N, F, S) are defined in Table 1.

Table 3

Examples of entries in the glycan structure database

No	Glycan structure	Symbol representation	Sequences in database
1		$\begin{array}{c} \text{H-N-H} \\ \diagdown \quad \diagup \\ \text{S-H-N-H} \quad \text{H-N-N-K} \end{array}$	HNHSHNHHNNK SHNHHNHHNNK HNSHNHHHHNNK SHNHNHHHHNNK
2		$\begin{array}{c} \text{N-H} \quad \text{N} \quad \text{F} \\ \diagdown \quad \diagup \quad \\ \text{N-H} \quad \text{H-N-N-K} \end{array}$	NHFNHNHHNNK NHNHNHNHFNK NNHHNHNHFNK NHNHNHFNHNNK NHNHFNHNNK
3		$\begin{array}{c} \text{S-N-H} \\ \diagdown \quad \diagup \\ \text{S-N-H} \quad \text{H} \\ \diagdown \quad \diagup \\ \text{S-H-N-H} \quad \text{H-N-N-K} \end{array}$	SHNSHNSHNNHHNNK SSHNNHNSHNNHHNNK SHNHSNHSNHHNNK SSHNNHNSHNNHHNNK

Table 4

Results of GlyDB search against a library of five isomeric structures with the same composition as the synthetic glycopeptide, see Fig. 3A, structure No. 1.

Structure ^a	Glycan sequences	Xcorr	Other matches ^b
1	NNHHNHNFNK	3.064	
1	NHNHNHNFNK	2.916	
1	FNNHHNHNK	0.995	2
1	FNHNHNHNK	0.937	1
3	NHNNHHNFNK	2.542	
3	NNHNHHNFNK	2.366	
3	NNNHHNHNFNK	1.786	
3	FNNHHNHNK	1.154	1
3	FNNHHHNK	0.961	2
4	NNHFNHNK	1.973	
4	NNFNHNK	1.417	
4	FNHNHNK	1.096	
5	NHFNHNK	1.727	
5	NFNHNK	1.279	
2	NHFNHNK	1.568	
2	NFNHNK	1.314	

^a Glycan structures shown in Fig. 3A.

^b Number of other glycans with the same linear sequence.

Table 5A
Summary of GlyDB searches for EGFR glycosylation site N-377

N	Glycan type	Glycan sequence	MH+	Charge	Xcorr
1	High Hex5 ^a	HHHHHNNK	5222.5	5	2.14
2	High Hex6 ^a	HHHHHHNNK	5384.7	4	1.59
		HHHHHHNNK	5384.7	5	1.75
3	High Hex7 ^a	HHHHHHHHNNK	5546.9	5	1.51
4	High Hex8 ^a	HHHHHHHHHHNNK	5709.1	4	1.80
		HHHHHHHHHHNNK	5709.1	5	1.75
5	Four-branches N-N-SHN-SHN-Fr ^b	NHNHSHNSHNHNFNK	6763.3	6	1.68

Other symbols as in Fig. 1.

^a Structures 1 to 4 most likely correspond to high mannose type glycans.
^b Fr — reducing terminal fucose.

Table 5B
Summary of GlycDB searches for EGFR glycosylation site N-480.

N	Glycan type	Glycan sequence	MH+	Charge	Xcorr
1	Two-branches	SHN-SHN			
		SSHNNHHHHNNK	4675.2	4	1.97
		SSHNNHHHHNNK	4675.2	4	2.28
		SHNHSHHHHHNNK	4675.2	4	2.29
		SHNSHHHHHHNNK	4675.2	4	2.77
2	Four-branches	SHN-FHN-SHN-FHNHN			
		HNSHNSHNFHNHHHHNNK	5917.1	5	1.52
3	Four-branches	FHN-FHN-SHN-SHNHN			
		HNFHNSHNSHNNHHHHNNK	5917.1	5	1.62
4	Two-branches	SHN-FHN-FP ^b			
		SHNHFHNNHNNFNK	4676.5	4	1.87