

Automated glycopeptide analysis—review of current state and future directions

David C. Dallas, William F. Martin, Serenus Hua and J. Bruce German

Submitted: 16th April 2012; Received (in revised form): 26th June 2012

Abstract

Glycosylation of proteins is involved in immune defense, cell–cell adhesion, cellular recognition and pathogen binding and is one of the most common and complex post-translational modifications. Science is still struggling to assign detailed mechanisms and functions to this form of conjugation. Even the structural analysis of glycoproteins—glycoproteomics—remains in its infancy due to the scarcity of high-throughput analytical platforms capable of determining glycopeptide composition and structure, especially platforms for complex biological mixtures. Glycopeptide composition and structure can be determined with high mass-accuracy mass spectrometry, particularly when combined with chromatographic separation, but the sheer volume of generated data necessitates computational software for interpretation. This review discusses the current state of glycopeptide assignment software—advances made to date and issues that remain to be addressed. The various software and algorithms developed so far provide important insights into glycoproteomics. However, there is currently no freely available software that can analyze spectral data in batch and unambiguously determine glycopeptide compositions for *N*- and *O*-linked glycopeptides from relevant biological sources such as human milk and serum. Few programs are capable of aiding in structural determination of the glycan component. To significantly advance the field of glycoproteomics, analytical software and algorithms are required that: (i) solve for both *N*- and *O*-linked glycopeptide compositions, structures and glycosites in biological mixtures; (ii) are high-throughput and process data in batches; (iii) can interpret mass spectral data from a variety of sources and (iv) are open source and freely available.

Keywords: *glycopeptide; glycoproteomics; glycopeptidomics; bioinformatics; N-linked; O-linked*

INTRODUCTION

Protein glycosylation, i.e. the enzymatic addition of carbohydrate chains to a protein, is estimated to occur on more than half of all eukaryotic proteins [1]. During synthesis, a set of more than 200 competing glycosyltransferases modify the nascent glycoprotein by adding specific monosaccharide structures via a specific linkage [2]. Therefore, protein glycosylation is a dynamic process—depending on its biochemical environment, a glycosylation site can be

differentially glycosylated with multiple glycan types, or completely unglycosylated [2]. A glycoprotein containing just four sites of glycosylation, with four possible glycans at each site, can theoretically have 625 different glycoforms (five possible glycosylation states, including unglycosylated for four sites: 5^4 combinations).

Glycosylation plays an intricate role in protein form and function. Glycosylation is involved in immune defense [3], cell–cell adhesion [4, 5], cellular

Corresponding authors. J. Bruce German, 1136 Robert Mondavi Institute, North, Department of Food Science and Technology, One Shields Avenue, University of California, Davis, Davis, CA, 95616. Tel: +530-752-1486; E-mail: jbgerman@ucdavis.edu

David Dallas is a Ph.D. student in Nutritional Biology at the University of California, Davis and specializes in mass spectrometric analysis of digestion of human milk proteins in infants.

William F. Martin is a Ph.D. student in Nutritional Biology at the University of California, Davis and specializes in programming and bioinformatics.

Serenus Hua is a Ph.D. student in Agricultural Chemistry at the University of California, Davis and specializes in mass spectrometric analysis of glycans and glycopeptides.

J. Bruce German is a professor in the Department of Food Science and Technology at the University of California, Davis and is Director of the Foods for Health Institute.

recognition [6] and pathogen binding [7]. Both compositional and structural changes in glycosylation are associated with a variety of disease states, such as breast cancer, ovarian cancer, prostate cancer, hepatocellular carcinoma and ocular rosacea [8–19].

In order to better understand the functions and bioactivities of glycoproteins and glycopeptides, researchers must develop methods to determine glycoprotein and glycopeptide composition and structure. However, from an analytical standpoint, isolation and analysis of such complex and heterogeneous molecules is extremely difficult, especially within their biological context [4, 20]. To simplify analysis, researchers often separate the glycans from the glycoprotein via chemical [21] or enzymatic [22, 23] reactions so that the identity of each component can be determined. Though this approach is less analytically complex than identification of intact glycopeptides, it prevents determination of specific sites of glycosylation (glycosites) as well as the identity of the glycosylated proteins.

Proteolytic cleavage of a glycoprotein breaks up the glycoprotein into multiple glycopeptides while preserving the glycan–protein linkages, which allows determination of glycan heterogeneity at specific sites on a particular glycoprotein [24–30]. Identification of protein-linked glycosylation in the context of the site of glycosylation provides insights into glycoprotein structure and function, making proteolytic cleavage of glycoproteins the preferred method for site-specific glycosylation analysis.

To analyze naturally occurring glycopeptides in a biological mixture, these glycopeptides must first be isolated. Methods for isolation of glycopeptides have been reviewed recently by Zhang *et al.* [31]. While enrichment of glycopeptides aids in sample analysis, clean-up is rarely complete. Therefore, an automated glycopeptide analysis program must avoid wrongly identifying such contaminants as glycopeptides.

Because of the difficulty in isolating and analyzing branched, heterogeneous molecules such as glycoproteins and glycopeptides, high mass-accuracy mass spectrometry (MS) stands out as the most promising platform for glycoprotein analysis. MS provides sufficient mass accuracy to differentiate between compounds only fractions of a Dalton apart in mass. Coupled with tandem fragmentation, which breaks down single molecules into smaller pieces, MS can be used to solve structure. MS coupled

with chromatographic separation allows structurally specific multi-dimensional analysis [32]. High-resolution MS platforms are used successfully for site-specific glycosylation analysis of glycoproteins and glycoprotein mixtures [24–30].

The high sensitivity and multi-dimensionality of MS-based glycosylation analysis platforms require that vast amounts of data be generated in order to fully capture the complexity of the information gathered. Processing these data in a high-throughput manner requires the development and adaptation of computational tools and algorithms capable of elucidating glycoprotein/glycopeptide compositions and structures. This review examines the capabilities of current glycopeptide analytical methods and delineates the issues that remain to be solved by future glycoproteomic analysis software.

GLYCOPROTEOMIC ANALYSIS SOFTWARE

Automated peptide identification employing both single and tandem MS data is now routine, and possible with several software packages [33, 34]. Most of these software packages can handle simple post-translational modifications such as phosphorylation and single sugar glycosylation; however, none are capable of searching for more complex glycans. In the past decade, numerous software packages were created for the assignment of free glycan compositions. These software packages include GlycoWorkbench [35], SysBioWare [36], GlycoFragment [37], SimGlycan [38], StrOligo [39], GlycoSearchMS [37] and Cartoonist [40]. Several software packages that employ tandem MS/MS data to determine glycan structure were created. These glycan structural assignment software packages include STAT [41], OSCAR [42], GLYCH [43], GlycosidIQ [44], SimGlycan [38] and GlycoWorkbench. Following the creation of these software packages, various research groups released software packages aimed at solving glycopeptide compositions.

Data from experiments analyzing glycopeptides are especially difficult to handle from a programming perspective because of the complexity and size of the molecules. Because of their large size, large number of monomer units (5 unique sugar masses and 19 unique amino acid masses) and complex structure, assigning glycopeptide composition to intact masses (molecular weights) yields multiple compositional

Table 1: Table of glycopeptide analysis software versus inputs

Software name	Uses intact mass to find possible glycopeptide compositions	Uses protein sequence	Uses biological glycosite filter	Incorporates allelic variation	Does not require previous knowledge of glycans	Does not require peptide mass input	Uses tandem data	Uses retention time
Branch-and-Bound	×	×	×	×	✓	✓	✓	×
GlycoMaster	×	×	×	×	✓	×	✓	×
GlycoMiner	✓	✓	✓	×	✓	✓	✓	×
GlycoMod	✓	✓	✓	×	✓	×	×	×
Glycopeakfinder	✓	×	×	×	✓	×	✓	×
GlycoPep DB	✓	✓	✓	×	✓	×	×	×
GlycoPep grader	✓	✓	×	×	×	×	✓	×
GlycoPep ID	×	✓	×	×	×	×	✓	×
GlycoSpectrumScan	✓	×	×	×	×	×	×	×
GlycoX	✓	✓	×	×	✓	×	×	×
GlyDB	✓	✓	✓	×	✓	×	✓	×
GlyPID	✓	×	unclear	×	✓	✓	✓	×
GP finder	✓	✓	✓	×	✓	✓	✓	×
Medicel N-glycopeptide library	✓	✓	✓	×	✓	✓	✓	×
Peptonist	✓	✓	✓	×	✓	✓	✓	×
Sweet substitute	×	×	×	×	✓	✓	✓	×
Ideal software	✓	✓	✓	✓	✓	✓	✓	✓

possibilities (even with low parts-per-million error range cut-offs). To determine which composition is correct, researchers typically isolate and fragment individual masses (tandem MS/MS) and then analyze the resulting fragments for composition. Though necessary, fragmentation increases the complexity of the data and makes for a more challenging analytical problem. Even for a skilled data interpreter, solving each tandem spectrum by hand takes time and is error prone. Attempting to solve a large number of glycopeptides isolated from a complex biological system becomes a prohibitively time-intensive project. Therefore, an automated approach is necessary for high-throughput glycopeptide determination.

To date, several research groups have released software packages for determining glycopeptide composition, and each group has made important steps toward a solution. The software packages include Peptonist [45], GlycoMod [46], GlycoMaster [47, 48], Medicel N-glycopeptide library [49], GlycoPep DB [50], GlycoPep ID [51], GlyPID [52], Branch-and-Bound [53], GlycoMiner [54], GlyDB [55], GlycoSpectrumScan [56], GlycoPeakfinder [57], Sweet Substitute [58], GlycoX [59], GP Finder [24, 28, 30] and GlycoPep Grader [60]. Tables 1–4 summarize the attributes of each software package.

CHALLENGES IN AUTOMATED GLYCOPEPTIDE ANALYSIS

Automated glycopeptide analysis is difficult for a variety of reasons. This review will explore the key challenges in the creation of an automated glycopeptide analysis software package.

Intact mass alone is not enough to solve glycopeptide composition

With high mass-accuracy MS, free glycans can be compositionally characterized solely on the basis of intact *m/z*. Glycopeptides cannot be identified by this methodology, however, because typically more than one composition is possible within the error window of each intact mass. Glycopeptide masses are more likely than free glycans to match multiple possible compositions within the error range of an intact mass because they can contain more unique mass monomer units and often have greater masses than glycans. Glycans can contain only a small number of monosaccharide masses [only four or five in human biology (hexoses (Hex), *N*-acetylhexosamines (HexNAc), fucose (Fuc) and *N*-acetyl and, rarely, *N*-glycolyl neuraminic acids (NeuAc, NeuGc))]. Glycopeptides, on the other hand, can be comprised of Hex, HexNAc, Fuc, NeuAc, NeuGc and all 19 uniquely massed amino acid components (leucine and isoleucine are isomers that have

Table 2: Table of glycopeptide analysis software versus data handling abilities

Software name	Performs internal deconvolution and deisotoping	Compatible with a variety of data input types	Handles data in batch
Branch-and-Bound	×	Unclear	×
GlycoMaster	×	Mass/intensity text files only	×
GlycoMiner	×	Input as ProteinLynx xml project (Waters), or spectra in ASCII peak list format (PKL-available on most instruments)	✓
GlycoMod	×	Mass/intensity text files only	×
Glycopeakfinder	×	Mass/intensity text files only	×
GlycoPep DB	×	Mass/intensity text files only	×
GlycoPep grader	×	Mass/intensity text files only	×
GlycoPep ID	×	Mass/intensity text files only	×
GlycoSpectrumScan	×	Mass/intensity text files only	×
GlycoX	×	Mass/intensity text files only	×
GlyDB	✓	Unclear	✓
GlyPID	✓	Unclear	Unclear
GP finder	×	Mass/intensity/rt text files only	×
Medicel N-glycopeptide library	×	Micromass Ltd. pkl file format only	✓
Peptonist	✓	✓	✓
Sweet substitute	×	Mass/intensity text files only	×
Ideal Software	✓	✓	✓

Table 3: Table of glycopeptide analysis software versus outputs

Software name	Solves peptide, glycan and glycosite of N- and O-linked glycopeptides	Solves glycopeptides from single protein, protein cocktails and complex biosystems	Determines glycan structure	Provides match score	Calculates false positive and false negative rate
Branch-and-Bound	×	✓	✓ (in some cases)	✓	×
GlycoMaster	×	×	✓	✓	×
GlycoMiner	×	×	×	✓	✓
GlycoMod	×	×	×	×	×
Glycopeakfinder	✓	×	×	×	×
GlycoPep DB	×	×	×	×	×
GlycoPep Grader	×	×	×	✓	×
GlycoPep ID	×	×	×	×	×
GlycoSpectrumScan	✓	×	×	×	×
GlycoX	✓	×	×	×	×
GlyDB	×	×	✓ (for simple biantennary glycans)	✓	×
GlyPID	×	×	×	✓	×
GP finder	✓	✓ (for pronase digested glycopeptides)	×	✓	×
Medicel N-glycopeptide library	×	✓	×	✓	×
Peptonist	×	×	×	✓	×
Sweet substitute	×	×	✓ (in some cases)	×	×
Ideal software	✓	✓	✓	✓	✓

the same mass). With more monomer units, there are more possible compositions with masses within the mass error range. As knowledge of an intact mass number is typically not sufficient to identify glycopeptides, researchers fragment individual molecules

to determine which compositional possibility is correct. Retention time can be used in addition to intact mass to determine glycopeptide composition, but only after the molecule is identified by tandem MS/MS in a previous experiment.

Table 4: Table of glycopeptide analysis software versus flexibility and availability

Software name	Works on all computer platforms	Freely available (made clear in article)
Branch-and-Bound	Unclear	×
GlycoMaster	Unclear	×
GlycoMiner	×(windows only)	✓(w3.chemres.hu/ms/Glycominer)
GlycoMod	✓ (web-based)	✓(expasy.org/tools/glycomod/)
Glycopeakfinder	✓ (web-based)	✓(glyco-peakfinder.org/)
GlycoPep DB	✓ (web-based)	✓(hexose.chem.ku.edu/sugar.php)
GlycoPep Grader	✓ (web-based)	✓(glycopro.chem.ku.edu/GPGHome.php)
GlycoPep ID	✓ (web-based)	✓(hexose.chem.ku.edu/sugar.php)
GlycoSpectrumScan	✓ (web-based)	✓ (glycospectrumscan.org)
GlycoX	×(windows only)	× (but available upon request from author ^a)
GlyDB	Unclear	×
GlyPID	×(windows only)	×
GP Finder	×(windows only)	× (but available upon request from author ^a)
Medicel N-glycopeptide library	Unclear	×
Peptoonist	Unclear	×
Sweet substitute	×(windows only)	× (but available upon request from author ^b)
Ideal software	✓	✓

Contact information for software available upon request: ^aDr. Carlito Lebrilla, cblebrilla@ucdavis.edu; ^bDr. Stefan Clerens, stefan.clerens@bio.kuleuven.ac.be.

Compositions must be assigned to tandem MS/MS fragments

Determining the composition of each fragment of a glycopeptide helps to determine the correct precursor molecule composition. Solving for fragment compositions increases the computational work for the analytical software. Isotope patterns (the distribution of the molecule in its various isotopic compositions) in tandem MS spectra are lower quality than those in single MS spectra, which makes assignment of the monoisotopic ion (the molecule with no additional neutrons) by an algorithm more difficult. If isotope patterns are clear enough, algorithms can be designed to assign charge state and select the monoisotope based on the isotope pattern. The unclear isotope data often found in tandem spectra can preclude charge state and monoisotope assignment. Researchers must determine how glycopeptide analytical software will resolve unclear isotope data.

Glycan structure must be solved

Without determination of glycan structure, researchers will only be able to scratch the surface of glycoprotein and glycopeptide biological function. Glycan structural variation is large and derives from linkage type (variations in connectivity at each glycosidic linkage), branching arrangement (the order in which monosaccharides are attached to one another) and anomericity (whether the

glycosidic bond is alpha or beta). Solving for structure with MS is more difficult than solving for composition. One technique for determining glycan structure via MS involves looking for fragments produced from the intact molecule via cross-ring cleavages (or glycan losses via interior bonds rather than glycosidic bonds) as they can reveal information about connectivity [61].

Existing glycan structural libraries can also be used to help determine protein-linked glycan structure. However, researchers must be careful in employing this approach for two reasons. First, not all glycan structures are known, and not all known glycans are in the databases; therefore, employing known glycan structures too rigidly will preclude discovery. Second, a single glycan composition can exist in multiple structural isomers, therefore, assigning structure based on compositional match to a known structure precludes new structural discovery for that glycan composition. Therefore, an ideal software package would use known glycan structural libraries to suggest plausible structures or lend support to software-assigned structures, but not to curtail the structural possibilities to previously discovered structures.

Various glycan structural families exist

The existence of two distinct glycan families—*N*- and *O*-linked glycans—complicates glycopeptide analysis further. These families vary both in amino

acid connectivity—*N*-linked glycans attach at asparagine residues in the consensus sequence Asn–Xxx–Thr/Ser [5, 62, 63], whereas *O*-linked glycans attach at serine or threonine residues [64]—and in core structures—*N*-linked glycans have a single pentasaccharide core structure [2], whereas *O*-linked have at least eight disaccharide core structures [64]. Any complete glycopeptide analysis software package must be capable of determining compositions of both classes of glycans.

Glycopeptides can be multiply glycosylated

Single peptides can have more than one site of glycosylation. That peptides can be multiply glycosylated further complicates glycopeptide analysis. Multiple glycosylation is especially common for *O*-linked glycopeptides, as *O*-linked glycans tend to occur in tight clusters in the amino acid sequence [65]. Glycopeptide analysis software must be able to recognize peptides as multiply glycosylated and not lump glycans together as a single glycan chain. Multiply glycosylated peptides occur more commonly when produced by hydrolysis with specific proteases than with Pronase—a commercially available mixture of proteinases isolated from *Streptomyces griseus*—because a specific enzyme's cleavage site can be effectively blocked by steric hindrance, whereas the cocktail of enzymes in Pronase limits this problem.

Glycopeptides can be multiply charged

Charges can be localized in both the glycan and peptide moieties of glycopeptides [27]. Because of this combination of potential charge sites, glycopeptides can ionize in a variety of charge states when electrospray ionization is employed in MS. Potential to exist in a variety of charge states adds complexity to the data in two ways. First, charge state must be determined via isotope pattern. Second, multiply charged precursor molecules give rise to tandem MS/MS fragments at a variety of charge states, each of which must be determined via its isotope pattern.

Not all glycopeptides will be fragmented in a single LC-MS/MS analysis

For glycopeptide analysis by LC-MS/MS, researchers typically use data-dependent acquisition for fragmentation based on ion abundance or charge state. Not all glycopeptides in a sample will

be selected by the instrument for fragmentation with this approach. Without fragmentation spectra, glycopeptide composition cannot be confirmed. High-throughput glycopeptide analysis software should, therefore, guide secondary LC-MS/MS analysis to collect fragmentation spectra for potential glycopeptide masses not fragmented in the first analysis. Only one glycopeptide analysis software package, GlyPID [52], helps to guide iterative sample analyses until all potential glycopeptides are successfully fragmented.

Multiple fragmentation energies may be required to gain enough information

Standard collision-induced dissociation (CID) fragmentation of glycopeptides typically reveals mostly losses of glycans via glycosidic bond cleavage, and relatively little peptide bond cleavage [27]. This lack of information means that determining the peptide identity will be difficult. A combination of a low energy fragmentation for glycan component identification and a high energy fragmentation for peptide degradation may be necessary to identify a compound. Thus, the ideal program would identify the compounds that require further fragment information and guide further experiments. Then, the program would have to be capable of combining the information from two or more energy levels for identification.

CAPACITY OF CURRENT GLYCOPEPTIDE ANALYSIS SOFTWARE TO SOLVE HIGH-THROUGHPUT GLYCOPROTEOMICS PROBLEMS

As can be seen from inspection of Tables 1–4, no currently available glycopeptide analysis software has all of the capabilities necessary for high-throughput glycoproteomics.

Input

A high-throughput glycopeptide analysis software package should assign all possible compositions of each intact mass within a specified error window. Several glycopeptide analysis packages, including Branch-and-Bound, GlycoMaster, GlycoPep ID and Sweet Substitute, lack this capability. Many programs are capable of assigning possible glycopeptide compositions from intact mass, namely GlycoMiner, GlycoMod, Glycopeakfinder, GlycoPep DB,

GlycoSpectrumScan, GlycoX, GlyDB, GlyPID, GP Finder, Mediceal N-glycopeptide library, Peptoonist and GlycoPep Grader (see Table 1).

Allowing only peptide segment possibilities that match the sequence of proteins known to exist in the biological system from which the sample is derived also narrows the search space for glycopeptide compositions. The glycopeptide's protein of origin can be determined when peptide mass information is matched to biologically relevant protein sequences. Several of the current analytical software packages, including Branch-and-Bound, GlycoMaster, Glyco-Peakfinder, GlyPID and Sweet Substitute, do not match peptides to known protein sequences; however, several programs do match the peptides to protein sequences, namely GlycoMiner, GlycoMod, GlycoPep DB, GlyDB, GP Finder, Mediceal N-glycopeptide library, Peptoonist and GlycoPep Grader (see Table 1).

Algorithms specifically designed to assign glycans only at sites known or predicted to be glycosylated also narrow the search space. Predicting algorithms can be based on a rule that glycosylation can occur only on the Asn in the Asn-Xxx-Thr/Ser/Cys consensus sequence (N-linked) or on Ser/Thr (O-linked). This glycosite filter will lower the number of false positive glycopeptide compositions. Branch-and-Bound, GlycoMaster, Glyco-Peakfinder, GlycoPep ID, Sweet Substitute and GlycoPep Grader lack this capability (see Table 1). The programs that employ a glycosite filter are GlycoMiner, GlycoMod, GlycoPep DB, GlyDB, GP Finder, Mediceal N-glycopeptide library and Peptoonist.

Finding all possible glycopeptides in a system necessitates taking into account allelic variation in protein sequences. If allelic variation is not accounted for, the software would be unable to assign compositions to any glycopeptides possessing variant alleles. No existing glycopeptide analysis software accounts for allelic variation (see Table 1).

For streamlined glycoproteomics discovery, programs should allow glycan/glycosite determination without multiple experimental procedures. For some existing glycopeptide software, users must first determine all N- and O-linked glycan compositions after chemical or enzymatic release. Therefore, for these programs, at least one additional experiment must be performed before glycosite and protein of origin determination are possible. Requiring additional experiments for identification should be

avoided. However, multiple experiments can aid in identification, so a glycopeptide program should allow for multiple experimental inputs but not require them. Two available glycopeptide analysis software packages—GlycoPep ID and GlycoSpectrumScan—require previous data on exact N- and O-glycan compositions (see Table 1).

High throughput glycopeptide identification ideally would not require the user to input the known peptide masses. Several software packages, including GlycoMaster, GlycoMod, Glyco-Peakfinder, GlycoPep DB, GlycoPep ID, GlycoSpectrumScan, GlycoX, GlyDB and GlycoPep Grader, require that the user first solve for the peptide mass of each glycopeptide before the software can assign a glycopeptide composition. Therefore, to use these programs, researchers must spend more time determining and inputting peptide masses before solving glycopeptides. Determination of peptide masses in a sample can be difficult—for example, samples prepared via Pronase digest create numerous peptide segments because of non-specific cleavage (see Table 1).

High-confidence glycopeptide assignment needs tandem fragmentation data. Researchers can use tandem fragment data to determine which of a number of possible compositions for a mass is correct. GlycoMod, GlycoPep DB, GlycoSpectrumScan and GlycoX do not make use of tandem MS/MS information (see Table 1). Tandem information is used, however, in Branch-and-Bound, GlycoMaster, GlycoMiner, Glycopeakfinder, GlycoPep ID, GlyDB, GlyPID, GP Finder, Mediceal N-glycopeptide library, Peptoonist, Sweet Substitute and GlycoPep Grader.

For decades, chromatographic retention time has been used to aid in molecular identification. In glycopeptide analysis, retention time can help to determine compound identity. The only glycopeptide program that incorporates retention time data is GP Finder (see Table 1). Retention time is only recorded in GP Finder however; this information is not then used to automatically identify glycopeptides found in following experiments.

Data handling

Deisotoping and deconvolution are essential for mass analysis. Deisotoping is the process of selecting the monoisotopic ion from within an isotopic cluster in mass spectra. Deconvolution is the process of determining the charge state of the monoisotope.

Differences in m/z values between ions in the isotope cluster are used to determine charge state. Some mass spectral analytical platforms do not provide batch deconvolution and deisotoping of both single and tandem mass spectral data. Therefore, glycopeptide software that works with all platforms must perform internal deconvolution and deisotoping. Several current software packages have this capability, but many do not, including Branch-and-Bound, GlycoMaster, GlycoMiner, GlycoMod, Glyco-Peakfinder, GlycoPep DB, GlycoPep ID, GlycoSpectrumScan, GlycoX, GP Finder, MediceL N-glycopeptide Library, Sweet Substitute and GlycoPep Grader (see Table 2).

In order to gain widespread acceptance in the glycoproteomics field, a glycopeptide analysis software package must be able to use mass spectral data produced from a wide variety of instruments. Some software like GlycoMiner and Peptonist are compatible with multiple data formats, but most software programs are limited in this respect (see Table 2). The majority of current glycopeptide analysis programs only allow input as mass/intensity text files (i.e. a simple file of mass in one column and intensity in the other either tab-delimited or comma-separated) and do not allow for the more common proteomic data file formats such as .mgf, .mzXML or .mzML (see Table 2).

A high-throughput glycopeptide analysis program must be capable of handling and analyzing data in batch. Users must be able to import entire LC-MS-MS/MS data sets to the software at once and the software must be able to analyze all of this data at once. Use of programs lacking batch import and analysis capabilities is too time-consuming for high-throughput data analysis. Most current glycopeptide analysis software packages lack batch import capabilities. Software packages lacking batch import and analysis capacity include Branch-and-Bound, GlycoMaster, GlycoMod, Glyco-Peakfinder, GlycoPep DB, GlycoPep ID, GlycoSpectrumScan, GlycoX, GP Finder, Sweet Substitute and GlycoPep Grader (see Table 2).

Ideally, the data flow should capture the full context of the biological molecule, including location in the protein sequence, allelic variations and non-glycan modifications. Some existing software operates on amino acid sequences, divorced from their biological context, which requires later manual re-association of the results with their original context by the user. Software packages without

the ability to maintain context such as protein sequence include Branch-and-Bound, GlycoMaster, Glyco-Peakfinder, GlycoSpectrumScan, GlyPID and Sweet Substitute.

Output

Researchers need high-throughput glycoproteomics programs to provide, at minimum, basic outputs including glycosite, peptide sequence and glycan for both *N*- and *O*-linked glycans. A high-throughput glycoproteomics analysis program must determine compositions of both *N*- and *O*-linked glycopeptides. Unfortunately, most available glycopeptide analysis software packages can determine *N*-linked, but not *O*-linked glycopeptide compositions (see Table 3).

A high-throughput glycopeptide analysis software package must be able to narrow the many possible glycopeptide compositions to a single glycopeptide composition, whether the sample is a single protein, a protein cocktail or a complex biological system. Unfortunately, only a very few glycopeptide software packages can solve glycopeptides from a protein cocktail or from complex biological systems. MediceL N-glycopeptide library, Branch-and-Bound and GP Finder are the only programs purported to work with complex biological systems (see Table 3); however, MediceL N-glycopeptide library solves only *N*-linked compositions, Branch-and-Bound lacks batch analysis capacity and GP Finder works only for small glycopeptides digested by pronase.

Predicted and confirmed glycan structures are also desirable data outputs. Only four current software packages determine glycan structure of glycopeptides—Branch-and-Bound, GlycoMaster, GlyDB and Sweet Substitute. Each software package has limited structural determination capacity. For example, Branch-and-Bound typically determines a variety of possible structures and cannot determine which is correct. GlyDB can solve structures only for simple bi-antennary *N*-linked glycans. However, techniques presented in these software packages are important steps forward in glycopeptide analysis and should be incorporated into future glycopeptide determination software (see Table 3).

To differentiate between the selected identification and the other possibilities, a program must provide a score for each possible composition. In addition, the program would need a statistically appropriate cut-off determining when the top score is

significantly different from the rest to be trusted as correct. Such scoring mechanisms are common in proteomics platforms such as X!Tandem [66] and Mascot [67]. GP Finder, GlyPID, Medical N-glycopeptide library, Peptonist, GlyDB, GlycoMiner, GlycoMaster, Branch-and-Bound and GlycoPep Grader provide scores for matches (see Table 3).

Acceptable glycopeptide analysis software must provide users with false positive and false negative rates associated with a given result set. Researchers need these rates to determine which compositional assignments can be trusted (see Table 3). Of the current glycopeptide analysis programs, only GlycoMiner calculates and displays false positive and negative rates. Future software packages for high-throughput glycoproteomics should build in false positive and negative analysis.

After glycopeptide composition identification, comparisons of various biological samples will be needed. As LC/MS/MS data are complex, automation of such comparisons will speed data interpretation. No currently available software package provides post-identification sample comparison functionality. Future glycopeptide analysis programs should provide sample results comparisons.

Flexibility and portability

Scientific software packages are often designed to run on only one particular platform. Therefore, only researchers with access to that particular platform can utilize the software package. This lack of portability narrows the number of people who can adopt the software and benefit from the ease with which software can be used. Several glycopeptide software packages, namely GlycoMiner, GlycoX, GlyPID, GP Finder and Sweet Substitute work only on one platform. Operation of the algorithms via a web-interface avoids portability problems. Web-interfaces are used in GlycoSpectrumScan, GlycoPep ID, GlycoPep DB, Glyco-Peakfinder, GlycoMod and GlycoPep Grader (see Table 4). To best serve a wide variety of research groups, the ideal software would have at least two types of user interfaces: either a web site or a program with a graphical user interface supported on Windows, Mac and Linux and a scriptable program.

To be adaptable to ever-changing scientific questions and data sets, new software should be open source. Open source software can be expanded by any number of groups. As scientific research builds

upon the work of others, this flexibility is crucial. Currently, no glycopeptide analysis software programs are open source.

Availability

Glycopeptide analysis software should be freely available to the public. Free availability will allow more researchers to access this important software than if it were only available upon request or for purchase. Many current software packages are available to the public via Internet site download. Others are available only upon request by email to the author. Others can only be purchased, including Medical N-glycopeptide Library and Peptonist (see Table 4).

UNIQUE ASPECTS OF GLYCOPEPTIDE ANALYSIS SOFTWARE PACKAGES FOR INCORPORATION INTO FUTURE SOFTWARE

Searching for low mass saccharide ions in fragment spectra to screen for glycopeptides

GlycoMiner screens for glycopeptides by searching fragment spectra for low mass mono-, di-, tri- and tetra-saccharide oxonium ions (e.g. HexNAc, HexHexNAc, HexHexNAcFuc, HexHexNAcFucNeuAc). This strategy aids in data reduction by selectively isolating potential glycopeptides. This screening strategy should be considered for future glycopeptide analysis software.

Searching for glycopeptide isotope patterns to screen for glycopeptides

Peptonist scans all isotope profiles in single MS spectra to select potential glycopeptides. Peptonist considers isotope profiles as potential glycopeptides if they match the theoretical spectra of an average glycopeptide isotope profile (assuming an atomic formula of 50% hydrogen, 30% carbon, 5% nitrogen and 15% oxygen atoms). This approach identifies potential glycopeptides prior to fragmentation. Builders of future software should consider incorporating this technique as part of a screening system in searching for glycopeptides.

Searching for peptide mass alone to limit problem domain

GlycoMiner and Sweet Substitute search for characteristic ion series of monosaccharide losses leading to a peptide +HexNAc or the 'naked' peptide to determine the mass of the peptide alone. This strategy is possible for *N*-linked glycans because of the sequential loss of the HexNAcs from the *N*-glycan core (–HexNAc –HexNAc) followed by no more monosaccharide losses can be searched for. Researchers have not yet applied this strategy to *O*-glycans, which would have a –Hex –HexNAc pattern. Identification of the peptide mass reduces the problem of glycopeptide assignment to solving for just glycan composition. Solving for the rest of the mass (intact mass minus peptide mass) reveals far fewer compositional possibilities, as, with free glycans, compositions can only contain four to five monosaccharide component masses and, therefore, fewer compositions within the mass error range.

Explicit guidance for employing multiple collision energies

Data suggest that low energy CID leads to mostly glycosidic bond cleavages, whereas high energy CID yields more peptide bond cleavages [24, 27]. Both glycan and peptide loss information can aid in compositional and structural assignment. Therefore, researchers can obtain more compositional and structural clues from the fragment spectra by analyzing samples in both low and high energy CID. Branch-and-Bound directs users to use both low and high collision energy and can combine the two types of fragmentation spectra for assigning glycopeptide composition.

Searching for monosaccharide losses to aid in composition and structure determination

Searching for monosaccharide losses between fragment ions can help establish glycopeptide composition and basic structure. GlycoMiner and Sweet Substitute employ this strategy. Monosaccharide loss assignment can be error prone, however, because mass differences found can be due to random chance. Builders of future glycopeptide analysis software should consider adopting this technique.

Solving for possible glycan moiety structures

Branch-and-Bound employs known biological restraints to aid in assigning possible *N*-glycan structures. This program assigns structural possibilities by iteratively filtering out structures that do not match glycopeptide fragment ions.

GlyDB also assigns possible glycan structures of glycopeptides, but employs a different technique. GlyDB uses linear sequences and a set of structures created with glycan biosynthetic rules using the peptide identification program Sequest to determine glycan structure. GlycoMaster solves for possible glycan structures with a heuristic algorithm and matching against a generated set of suboptimal subtrees.

CONCLUSIONS AND PERSPECTIVES

In order to better understand the functions and bioactivities of glycoproteins, researchers must develop analytical tools capable of elucidating glycopeptide compositions and/or structures. Based on an analysis of current glycopeptide analytical software packages, the components and capabilities that should be included in future glycopeptide analysis software for high-throughput glycoproteomics are presented in Table 5.

Analysis of complicated glycopeptide data from complex biological systems such as human milk necessitates the construction of capable software packages. Until glycopeptides can be identified automatically, glycoproteomics will not catch up to the more mature 'omic' disciplines, namely proteomics, genomics, transcriptomics and metabolomics.

Determining glycoprotein structure and function remains one of the most challenging problems in proteomics [68]. Creating a program for high throughput glycopeptide compositional and structural analysis will aid the determination of protein-linked glycan functions. Toolsets for glycopeptidomics will accelerate advances in understanding biological structure-function relationships rivaling those of proteomics.

Open source

One of the major problems with scientific programs in general is that they are typically not open source. When software is available publicly but not open source, scientists can make use of the program, but

Table 5: Required components for high-throughput glycopeptide analysis programs

Programming areas	Requirements
Input	<ul style="list-style-type: none"> Match experimental intact masses to possible glycopeptides Employs biologically relevant protein sequences Provides biological filter for glycosite determination Incorporates allelic variation Does not require input of known glycans or peptide sequences Employs tandem MS/MS data Employs retention time
Data handling	<ul style="list-style-type: none"> Performs internal deconvoluting and deisotoping Compatible with multiple data input types Performs batch processing Retains biological context, such as position in protein sequence
Output	<ul style="list-style-type: none"> Solves peptide and glycan composition and glycosite for <i>N</i>- and <i>O</i>-linked glycopeptides Solves glycopeptide compositions in single protein, cocktail and complex biological samples Determines glycan structure Has low false positive and false negative rates Provides preliminary comparison of results from multiple samples
Flexibility	<ul style="list-style-type: none"> Provides options for different enzymatic cleavage types Open source
Portability	<ul style="list-style-type: none"> Works on Linux, Mac and Windows or web-interface
Availability	<ul style="list-style-type: none"> Freely available

only as far as the program's limits. Whenever a research question requires functionality beyond that provided by the program, researchers cannot make alterations to the program and, therefore, must rebuild the program entirely. Open source code allows researchers to adapt programs to new research questions and expanded functionality. For example, if the programs discussed in this article were open source, a programmer could build upon any one of them to extend functionality to structural determination, without having to rebuild the entire program. If a program only solved for *N*-linked glycopeptides, another programmer could modify the source code to solve *O*-linked glycopeptides as well. If one program had a successful algorithm for deconvolution and deisotoping, and another had an algorithm for matching possible compositions to single mass values, these two could be copied from the source code and incorporated into a novel program.

Open source programs are rare in science because programmers want to prevent other groups from employing their code for their own publications or because they are interested in privatizing the program for profit. As coding requires a large input of time, programmers want to publish numerous studies that utilize each program before the source code is made public. Problematically, however, by the time the authors have used the program in studies that

they publish, they often do not follow through with publishing the source code.

A summary of program function is often provided in a publication as information for other researchers to build upon. This summary, though important as an overview, is of little use from a programming perspective, as conceptually simple functions may be carried out from a bastion of hundreds of lines of precise coding. Replication of a code or a programmatic task based on the brief outline in publications is nearly impossible.

To fix this problem, publishers should require that articles demonstrating the function of a novel in-house program also include the source code as a supplement. Scientists themselves should push for publishing all programs open source. Moreover, the scientific community should build a virtual warehouse for scientific coding that can be sampled and assembled for novel programming by anyone. All newly published code should be automatically incorporated into this code repository. Such a resource would allow incredible advancements in science, especially in systems biology.

The lack of open source program publishing is one of the biggest problems in scientific programming today. If all glycopeptide analysis programs and other programs such as oligosaccharide and peptide analysis programs were made open source and freely available, the stumbling blocks of glycopeptide

analysis could be quickly overcome. When open source publishing of scientific programs becomes standard, the result will be enormous leaps in scientific progress, especially in high-throughput biological fields.

Key points

- High-throughput glycoproteomics and glycopeptidomics are not currently possible due to a lack of programs for analysis of single and tandem mass spectrometric data.
- No current glycoprotein analysis program provides all the functionality required for high-throughput glycoproteomics.
- Future programs should be open-source, freely available, web-based and be capable of using both MS and MS/MS information to identify both *N*- and *O*-linked glycopeptides from complex biological samples.

FUNDING

This work was supported by the National Science Foundation Graduate Student Research Fellowship; the National Institutes of Health Training Program in Biomolecular Technology [2-T3-GM08799]; the Jastro Shields Research Scholarship Award and the Graduate Scholars Fellowship.

References

1. 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**: 4–8.
2. Stanley P, Schachter H, Taniguchi N. *N*-glycans. In: Varki A, (ed). *Essentials of Glycobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2009.
3. Rudd PM, Elliott T, Cresswell P, et al. Glycosylation and the immune system. *Science* 2001;**291**:2370–6.
4. Dwek R. Glycobiology: toward understanding the function of sugars. *Chem Rev* 1996;**96**:683–720.
5. Varki A. *Essentials of Glycobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 1999.
6. Hakomori S. Carbohydrate-to-carbohydrate interaction, through glycosynapse, as a basis of cell recognition and membrane organization. *Glycoconj J* 2004;**21**:125–37.
7. Newburg D. Human milk glycoconjugates that inhibit pathogens. *Curr Med Chem* 1999;**6**:117–28.
8. An H, Ninonuevo M, Aguilan J, et al. Glycomics analyses of tear fluid for the diagnostic detection of ocular rosacea. *J Proteome Res* 2005;**4**:1981–7.
9. Kirmiz C, Li B, An HJ, et al. A serum glycomics approach to breast cancer biomarkers. *Mol Cell Proteomics* 2007;**6**:43–55.
10. de Leoz MLA, Young LJT, An HJ, et al. High-mannose glycans are elevated during breast cancer progression. *Mol Cell Proteomics* 2011;**10**:1–9.
11. An HJ, Miyamoto S, Lancaster KS, et al. Profiling of glycans in serum for the discovery of potential biomarkers for ovarian cancer. *J Proteome Res* 2006;**5**:1626–35.
12. Leiserowitz GS, Lebrilla C, Miyamoto S, et al. Glycomics analysis of serum: a potential new biomarker for ovarian cancer? *Int J Gynecol Cancer* 2008;**18**:470–5.
13. de Leoz MLA, An HJ, Kronewitter S, et al. Glycomic approach for potential biomarkers on prostate cancer: profiling of *N*-linked glycans in human sera and pRNS cell lines. *Dis Markers* 2008;**25**:243–58.
14. Bereman MS, Young DD, Deiters A, et al. Development of a robust and high throughput method for profiling *N*-linked glycans derived from plasma glycoproteins by nanoLC-FTICR mass spectrometry. *J Proteome Res* 2009;**8**: 3764–70.
15. Bereman MS, Williams TI, Muddiman DC. Development of a nanoLC LTQ Orbitrap mass spectrometric method for profiling glycans derived from plasma from healthy, benign tumor control, and epithelial ovarian cancer patients. *Anal Chem* 2009;**81**:1130–6.
16. Tang ZQ, Varghese RS, Bekesova S, et al. Identification of *N*-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data. *J Proteome Res* 2010;**9**: 104–12.
17. Hua S, An HJ, Ozcan S, et al. Comprehensive native glycan profiling with isomer separation and quantitation for the discovery of cancer biomarkers. *Analyst* 2011;**136**: 3663–71.
18. Zhao J, Simeone DM, Heidt D, et al. Comparative serum glycoproteomics using lectin selected sialic acid glycoproteins with mass spectrometric analysis: application to pancreatic cancer serum. *J Proteome Res* 2006;**5**:1792–802.
19. Reggi M, Capon C, Gharib B, et al. The glycan moiety of human pancreatic lithostathine. *Eur J Biochem* 1995;**230**: 503–10.
20. Brooks S, Dwek M, Schumacher U. *Functional and Molecular Glycobiology*. Oxford: BIOS Scientific Publishers Limited, 2002.
21. Greis KD, Hayes BK, Comer FI, et al. Selective detection and site-analysis of *O*-GlcNAc-modified glycopeptides by [beta]-elimination and tandem electrospray mass spectrometry. *Anal Biochem* 1996;**234**:38–49.
22. Tarentino AL, Gomez CM, Plummer TH Jr. Deglycosylation of asparagine-linked glycans by peptide: *N*-glycosidase F. *Biochemistry* 1985;**24**:4665–71.
23. Zhang W, Wang H, Zhang L, et al. Large-scale assignment of *N*-glycosylation sites using complementary enzymatic deglycosylation. *Talanta* 2011;**85**:499–505.
24. Hua S, Nwosu CC, Strum JS, et al. Site-specific protein glycosylation analysis with glycan isomer differentiation. *Anal Bioanal Chem* 2011;**403**:1–12.
25. An H, Peavy T, Hedrick J, et al. Determination of *N*-glycosylation sites and site heterogeneity in glycoproteins. *Anal Chem* 2003;**75**:5628–37.
26. Clowers B, Dodds E, Seipert R, et al. Site determination of protein glycosylation based on digestion with immobilized nonspecific proteases and Fourier transform ion cyclotron resonance mass spectrometry. *J Proteome Res* 2007;**6**: 4032–40.
27. Seipert RR, Dodds ED, Clowers BH, et al. Factors that influence fragmentation behavior of *N*-linked glycopeptide ions. *Anal Chem* 2008;**80**:3684–92.

28. Nwosu CC, Seipert RR, Strum JS, *et al.* Simultaneous and extensive site-specific N- and O-glycosylation analysis in protein mixtures. *J Proteome Res* 2011;**10**:2612–24.
29. Dodds ED, Seipert RR, Clowers BH, *et al.* Analytical performance of immobilized pronase for glycopeptide footprinting and implications for surpassing reductionist glycoproteomics. *J Proteome Res* 2008;**8**:502–12.
30. Nwosu CC, Strum JS, An HJ, *et al.* Enhanced detection and identification of glycopeptides in negative ion mode mass spectrometry. *Anal Chem* 2010;**82**:9654–62.
31. Zhang L, Lu H, Yang P. Specific enrichment methods for glycoproteome research. *Anal Bioanal Chem* 2010;**396**:199–203.
32. Hua S, Lebrilla C, An HJ. Application of nano-LC-based glycomics towards biomarker discovery. *Bioanalysis* 2011;**3**:2573–85.
33. Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;**20**:3551–67.
34. Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;**5**:976–89.
35. Ceroni A, Maass K, Geyer H, *et al.* GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J Proteome Res* 2008;**7**:1650–9.
36. Vakhrushev SY, Dadimov D, Peter-Katalinic J. *SysBioWare: Structure Assignment Tool for Automated Glycomics*. Potsdam: Beilstein-Institut, 2009.
37. Lohmann KK, Von Der Lieth CW. GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res* 2004;**32**:W261–6.
38. Albanese J, Glueckmann M, Lenz C. SimGlycan™ Software*: a new predictive carbohydrate analysis tool for MS/MS data. *Appl Biosystems* 2010.
39. Ethier M, Saba JA, Spearman M, *et al.* Application of the StrOligo algorithm for the automated structure assignment of complex N linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003;**17**:2713–20.
40. Goldberg D, Sutton Smith M, Paulson J, *et al.* Automatic annotation of matrix assisted laser desorption/ionization N glycan spectra. *Proteomics* 2005;**5**:865–75.
41. Sara P, Morrow J, Leary JA. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal Chem* 2000;**72**:2331–6.
42. Lapadula AJ, Hatcher PJ, Hanneman AJ, *et al.* Congruent strategies for carbohydrate sequencing. 3. OSCAR: An algorithm for assigning oligosaccharide topology from MS n data. *Anal Chem* 2005;**77**:6271–9.
43. Tang H, Mechref Y, Novotny MV. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* 2005;**21**:i431–9.
44. Joshi HJ, Harrison MJ, Schulz BL, *et al.* Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* 2004;**4**:1650–64.
45. Goldberg D, Bern M, Parry S, *et al.* Automated N-glycopeptide identification using a combination of single- and tandem-MS. *J Proteome Res* 2007;**6**:3995–4005.
46. Cooper CA, Gasteiger E, Packer NH. GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* 2001;**1**:340–9.
47. Shan B, Zhang K, Ma B, *et al.* GlycoMaster-A software for interpretation of glycopeptides from MS/MS spectra. *52nd ASMS Conference on Mass Spectrometry and Allied Topics*. 2004. Nashville, TN, USA.
48. Shan B. Stochastic context-free graph grammars for glycoprotein modelling. In: Domaratzki M, Okhotin A, Salomaa K, *et al.* (eds). 2005. Berlin/Heidelberg: Springer, 247–58.
49. Joenväärä S, Ritamo I, Peltoniemi H, *et al.* N-Glycoproteomics—An automated workflow approach. *Glycobiology* 2008;**18**:339–49.
50. Go EP, Rebecchi KR, Dalpathado DS, *et al.* GlycoPep DB: A tool for glycopeptide analysis using a “Smart Search”. *Anal Chem* 2007;**79**:1708–13.
51. Irungu J, Go EP, Dalpathado DS, *et al.* Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID. *Anal Chem* 2007;**79**:3065–74.
52. Wu Y, Mechref Y, Klouckova I, *et al.* Mapping site specific protein N glycosylations through liquid chromatography/mass spectrometry and targeted tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2010;**24**:965–72.
53. Peltoniemi H, Joenväärä S, Renkonen R. De novo glycan structure search with the CID MS/MS spectra of native N-glycopeptides. *Glycobiology* 2009;**19**:707–14.
54. Ozohanic O, Krenyacz J, Ludányi K, *et al.* GlycoMiner: a new software tool to elucidate glycopeptide composition. *Rapid Commun Mass Spectrom* 2008;**22**:3245–54.
55. Ren JM, Rejtar T, Li L, *et al.* N-Glycan structure annotation of glycopeptides using a linearized glycan structure database (GlyDB). *J Proteome Res* 2007;**6**:3162–73.
56. Deshpande N, Jensen PH, Packer NH, *et al.* GlycoSpectrumScan: fishing glycopeptides from MS spectra of protease digests of human colostrum sIgA. *J Proteome Res* 2010;**9**:1063–75.
57. Maass K, Ranzinger R, Geyer H, *et al.* “Glyco peakfinder”—de novo composition analysis of glycoconjugates. *Proteomics* 2007;**7**:4435–44.
58. Clerens S, Van den Ende W, Verhaert P, *et al.* Sweet Substitute: a software tool for in silico fragmentation of peptide linked N glycans. *Proteomics* 2004;**4**:629–32.
59. An HJ, Tillinghast JS, Woodruff DL, *et al.* A new computer program (GlycoX) to determine simultaneously the glycosylation sites and oligosaccharide heterogeneity of glycoproteins. *J Proteome Res* 2006;**5**:2800–8.
60. Woodin CL, Hua D, Maxon M, *et al.* GlycoPep Grader: A web-based utility for assigning the composition of N-linked glycopeptides. *Anal Chem* 2012;**84**:4821–4829.
61. Harvey DJ. Collision induced fragmentation of underivatized N linked carbohydrates ionized by electrospray. *J Mass Spectrom* 2000;**35**:1178–90.
62. Butler M, Quelhas D, Critchley AJ, *et al.* Detailed glycan analysis of serum glycoproteins of patients with congenital disorders of glycosylation indicates the specific defective glycan processing step and provides an insight into pathogenesis. *Glycobiology* 2003;**13**:601–22.
63. Ben-Dor S, Esterman N, Rubin E, *et al.* Biases and complex patterns in the residues flanking protein N-glycosylation sites. *Glycobiology* 2004;**14**:95–101.

64. Steen PV, Rudd PM, Dwek RA, *et al.* Concepts and principles of O-linked glycosylation. *Crit Rev Biochem Mol Biol* 1998;**33**:151–208.
65. Dube DH, Prescher JA, Quang CN, *et al.* Probing mucin-type O-linked glycosylation in living animals. *Proc Natl Acad Sci USA* 2006;**103**:4819–24.
66. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;**20**:1466–7.
67. Perkins DN, Pappin DJC, Creasy DM, *et al.* Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;**20**:3551–67.
68. Dell A, Morris HR. Glycoprotein structure determination by mass spectrometry. *Science* 2001;**291**:2351–6.