REVIEW

Glycobioinformatics: Current strategies and tools for data mining in MS-based glycoproteomics

Feng Li[1,2], Olga V. Glinskii[1,3] and Vladislav V. Glinsky[1,2]

[1] Research Service, Harry S. Truman Memorial Veterans Hospital, Columbia, MO, USA
[2] Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, MO, USA
[3] Department of Medical Pharmacology and Physiology, University of Missouri, Columbia, MO, USA

Glycobioinformatics is a rapidly developing field providing a vital support for MS-based glycoproteomics research. Recent advances in MS greatly increased technological capabilities for high throughput glycopeptide analysis. However, interpreting MS output, in terms of identifying glycan structures, attachment sites and glycosylation linkages still presents multiple challenges. Here, we discuss current strategies used in MS-based glycoproteomics and bioinformatics tools available for MS-based glycopeptide and glycan analysis. We also provide a brief overview of recent efforts in glycobioinformatics such as the new initiative UniCarbKB directed toward developing more comprehensive and unified glycobioinformatics platforms. With regards to glycobioinformatics tools and applications, we do not express our personal preferences or biases, but rather focus on providing a concise description of main features and functionalities of each application with the goal of assisting readers in making their own choices and identifying and locating glycobioinformatics tools most suitable for achieving their experimental objectives.

Received: April 10, 2012
Revised: October 6, 2012
Accepted: November 6, 2012

Keywords:
Bioinformatics / Glycan analysis / Glycomics / Glycoproteomics / MS

PTMs of proteins are the primary means used by prokaryotic and eukaryotic cells to regulate the activity of key proteins [1–3]. PTMs may involve both chemical alterations of protein side chains and a cleavage of the main chain peptide bonds. The dynamic modification and diversification enabled by PTMs greatly increases molecular variants of cellular proteins by an estimated one or two orders of magnitude over the number encoded by the genome [4]. Therefore, characterizing structures, sites, and dynamics of the protein PTMs is essential for understanding their diversity, structure, and function in the "-omics" age [5].

Correspondence: Dr. Vladislav V. Glinsky, M263 Medical Sciences Bldg., Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, MO 65212, USA
E-mail: glinskiivl@missouri.edu
Fax: +1-573-814-6551

Abbreviations: CFG, consortium for functional glycomics; **ECD**, electron capture dissociation; **ETD**, electron transfer dissociation; **IGAP**, intact glycopeptide analysis pipeline

1 Glycosylation: Most abundant and structurally diverse posttranslational modification

Covalent attachment of sugars or glycans to proteins or lipids is defined as glycosylation. Glycosylation represents not only the most abundant protein PTM, but also by far the most structurally diverse one [6]. Although monosaccharides do not constitute overly complicated chemical group, 13 different monosaccharides and eight different amino acids involved in glycoprotein linkages results in a total of at least 41 different glycan-protein bonds. Comprehensive information about various monosaccharide residues can be found in the MonosaccharideDB database (http://www.monosaccharidedb.org), which is being developed as part of the EUROCarbDB and GLYCOSCIENCES.de projects [7]. Furthermore, due to additional modifications on terminal glycans of oligosaccharide branches such as fucosylation, sulfation, acetylation, and/or sialylation (about 50 different sialic acids are known), the molecular diversity of

Colour Online: See the article online to view Figs. 1–3 in colour.

glycosylation rapidly increases exponentially and becomes incredibly complicated [8–10]. There are estimated 250–500 "glycogenes" in human genome (about 1–2% of the total genome), which are directly involved in glycan assembly [11]. Glycogenes are involved in the attachment and subsequent processing of the sugar portion of glycoconjugates and can be divided into several categories including glycosyltransferases, glycolytic enzymes, sugar nucleotide synthetases, and sugar nucleotide transporters, while glycosyltransferases comprise one of the largest and most diverse groups of enzymes with over 180 glycogenes being cloned and characterized to date [12, 13].

Consequently, it has been estimated that more than 50% of the entire human proteome is covalently modified with glycans, although latest reports suggest that less than one-fifth of proteins are glycosylated [14, 15]. In addition to a diversity of the attached glycan structures, macroheterogeneity (variable occupancy of several glycosylation sites), and microheterogeneity (variable degree of type, trimming, and elongation of the glycan attached to a single glycosylation site) contribute further to the complexity of protein glycosylation [16]. The macroheterogeneity and microheterogeneity of the glycosylation are controlled by multiple factors. For example, within a particular cell, not only the primary and secondary structure of the protein affects the location of the glycosylation sites and the level of their occupancy, but also the tertiary and quaternary structures influence the subsequent processing of the attached glycans [17]. Recent precision mapping of the in vivo *N*-glycoproteome clearly reveals rigid topological and sequence constraints of *N*-linked glycosylation of glycoproteins [18]. Thus, glycosylation is one of the most common PTMs existing in nature, which is characterized by the extreme structural diversity. Glycosylation plays essential role in many biological processes such as cell recognition, cell–cell communication, signaling, embryo development, immunity, etc. [19,20]. Therefore, identifying the glycosylation sites, their occupancy, and the attached glycan structures is crucial for proper understanding of the glycoprotein biological function. Yet tools for systematic identification and analysis of protein glycosylation are greatly underdeveloped.

2 Glycoproteomics: Systematic identification of glycosylation on proteome level

Based on the concept of proteome, the complete subset of glycosylated proteins (glycoproteins) generated by a cell or an organism under specific conditions, is defined as "glycoproteome." Therefore, the term "glycoproteomics" refers to studies that aim to define or quantify the complete set of proteins containing glycosylation modifications in a cell, tissue, or organism [19]. Rather than separating glycan and protein analyses into glycomics and proteomics, major tasks of glycoproteomics include not only identifying the protein main chains modified by glycosylation, but also assigning and/or mapping the structures and sites of these modifications [21]. Compared with identifying protein main chains, assigning or mapping glycan structures and sites in glycoproteomics is much more challenging. First, the biosynthesis of glycans is a nontemplate driven process involving coordinated expression of several glycosyltransferases, some of which have additional tissue-specific isoforms [22]. Consequently, the inherent heterogeneity and large diversity of glycan structures cannot be predicted from any reference database due to a complex biosynthesis and lack of proofreading machinery [23], although *N*-linked glycoproteins do have quality control "proofreading," whereby misfolded proteins or those proteins attached with $Glc1–3Man_9NAc_2$ are typically recycled with the help of calreticulin and calnexin in the ER [24, 25]. Second, based on the backbone chemical structure of glycosylation, glycans can be classified broadly as linear and branched sugars. Due to the branched nature of glycosylation, the chemical heterogeneity, and diversity of glycans challenge the development of analytical techniques such as MS to accurately define their chemical structures. Some glycosylation PTMs could be very difficult (if not impossible) to differentiate with MS technology alone, when individual monosaccharides having the same masses are involved, or the differences between glycans are caused not by the attachment of distinct chemical groups but by the different glycan-protein linkages or glycosidic linkages between individual monosaccharides are employed such as in core 5 *O*-glycans GalNAcα1–3GalNAcα-Ser/Thr and core 7 *O*-glycans GalNAcα1–6GalNAcα-Ser/Thr, or in core 1 *O*-glycans Galβ1–3GalNAcαSer/Thr, and core 8 *O*-glycans Galα1–3GalNAcαSer/Thr. In addition, the presence of heterogeneous mixtures of different chemical structures within glycans always need to be considered [26].

Based on the mode of attachment, two types of protein glycosylation with important biological functions are most common in nature: "*N*-linked" glycosylation (*N*-glycosylation), where glycans are attached to asparagines in a consensus sequence N-X-S/T (where X can be any amino acid except proline) via an *N*-acetylglucosamine (*N*-GlcNAc) residue; and "*O*-linked" glycosylation (*O*-glycosylation), where glycans are attached to serine or threonine through acyl linkages. *N*-Linked glycosylation is initiated via en bloc transfer of a tetradecasaccharide ($Glc_3Man_9GlcNAc_2$) from the lipid-linked oligosaccharide $Glc_3Man_9GlcNAc_2$-PP-dolichol to specific asparagine residues of nascent proteins in the ER by oligosaccharyltransferases [27]. Subsequent trimming or further elaboration to form one of the three standard types of *N*-glycan cores (high mannose, complex, or hybrid) is mediated by a series of glycosidases and glycosyltransferases in the Golgi apparatus based on this tetradecasaccharide [28]. *O*-linked glycans are formed by a series of glycosyltransferase-catalyzed steps that begin with the transfer of the first glycan from UDP-GalNAc directly to either a Serine or Threonine residue by one of a large number of polypeptide GalNAc-transferases found in the Golgi [29]. Different *O*-glycan core structures are synthesized by subsequent glycan addition based on the first glycan *O*-GalNAc (there are eight O-GalNAc-based glycan core

structures, most of which may be further glycosylated [30]. *N*- and *O*-glycosylation classification leads to further subdivision of glycoproteomics into *N*-linked and *O*-linked glycoproteomics focusing on these two important biologically functional PTMs.

Of note, it is important to remember that *O*-GlcNAc modification or *O*-GlcNAcylation is in many ways distinct from "classical" N- and O-linked glycosylation. *O*-GlcNAcylation is found mostly within the cytoplasm or nucleoplasm (with a few exceptions) and it is not elongated or further processed into a complex oligosaccharide [31]. *O*-GlcNAcylation is similar to protein phosphorylation, *O*-GlcNAc can be attached by *O*–GlcNAc transferase or removed by *O*–GlcNAcase, and this process is a dynamical response to changes in the cellular environment triggered by stress, hormones, or nutrients [32]. *O*-GlcNAcylation is very labile upon ionization in a mass spectrometer, and much of *O*-GlcNAc is often lost at the source, which makes *O*-GlcNAcylation very difficult to detect and map corresponding attachment sites [33]. In addition to *N*- and *O*-linked glycosylation, proteins can cross-link with reducing sugars and form advanced glycation end-products in a process called glycation or nonenzymatic glycosylation [34]. Strategies and tools for analyzing other known but less common forms of glycosylation such as C- and S-linked as well as other will not be discussed here.

3 MS technology: Method of choice in glycoproteomics

The general workflow in glycoproteomics consists of glycoprotein or glycopeptide enrichment, multidimensional protein or peptide separation, tandem mass spectrometric analysis, and bioinformatic data interpretation [35, 36]. Based on the general pipeline of glycoproteomics, two complementary strategies (the "bottom-up" and the "top-down") are currently widely used to identify proteins in glycoproteomics (Fig. 1) and each strategy has its own strengths and weaknesses (although sometimes a combination of these two complementary strategies being employed) [37–39]. In the "bottom-up" approach, the peptides resulting from proteolytic digestion or chemical cleavage of proteins are used for identification of peptide sequences and PTMs. Because the "bottom-up" strategy is based on peptide-centric approach, some PTMs may be ultimately unobtainable in MS as only a portion of the entire protein is generally detected. Further, PTMs and proteins resulting from alternative splicing as well as the same enzymatic peptide sequences from highly related protein families are difficult to be probed completely by the "bottom-up" approach. Nonetheless, "bottom-up" strategy plays a dominant role in the current glycoproteomics research, because it is more suitable for automation and high sample throughput making it easier to use employing current MS technologies [40, 41].

The advantage of the "top-down" approach is that sequence information and PTM data are acquired from intact proteins allowing for identification of protein isoforms, alternatively spliced isoforms, and PTMs once full-length protein sequence coverage has been demonstrated. The most significant recent advancements in the "top-down" approach are electron capture dissociation (ECD)and electron transfer dissociation (ETD) MS techniques, which are discussed in more detail later in this chapter. However, the "top-down" strategy currently still suffers from limited sensitivity and throughput.

No matter whether "bottom-up," "top-down," or a combination of the two approaches is used, MS is the method of choice for rapidly identifying the protein main chains and the structures and the sites of glycan attachment [42]. To elucidate peptide sequences, glycan structures, and the sites of their attachment on intact glycopeptides, they need to be isolated and dissociated (typically by a gas phase reaction) into smaller product ions in a MS instrument so as the product ions could be subsequently isolated and subjected to further dissociation reaction and mass analysis [43]. Thus, the control of the dissociation process from parent ions into product ions is the key aspect of MS/MS technique in glycopeptide analysis. Currently, CID, ECD, ETD, and infrared multi photon dissociation are the methods most widely employed to identify intact glycopeptides using MS/MS dissociation techniques [44]. CID, also known as collision activated dissociation, is a common fragmentation method utilizing a vibrational activation fragmentation process that breaks the weak bonds within peptides. However, if used to analyze *O*-linked glycopeptides, CID will normally return no or very low intensity information on the sites of *O*-linked modifications. Because *O*-linkages between glycans and serines or threonines are considerably more labile than peptide bonds, glycan residues will be generally eliminated during CID before peptide fragmentation [5, 45]. On the other hand, N-linkages are relatively more stable than *O*-linkages. Consequently, *N*-linked modifications have been successfully analyzed using CID, including identification of the peptide backbones and the sites of *N*-linked glycosylation [46].

To overcome these limitations, ECD and ETD dissociation techniques were developed to complement CID in MS analysis of *O*-glycopeptides and/or *N*-glycopeptides [47, 48]. ECD is a "mild" fragmentation technique based on partial recombination of multiply protonated polypeptide molecules with low-energy electrons (<0.2 eV), so the intact glycopeptides can be recorded by mass analyzer [49, 50]. However, the need of FT-ICR-MS greatly limited ECD use, though it has been reported on an ion trap MS as well [51]. ETD is a recently developed dissociation technique that shows promising alternative fragmentation pathways, which fragments peptides by transferring an electron from a radical anion (e.g. fluoranthene) to a protonated peptide [52]. Analogous to ECD, ETD results preferentially in the cleavages of the N–Cα bonds of the peptide backbone to generate homolog series of c'- and z'- type fragment ions without loss of the glycan moiety (Fig. 2). Compared to CID, ETD preserves glycosylation PTMs, which are often removed by CID, while sequence information being obtained for peptide identification [52, 53].

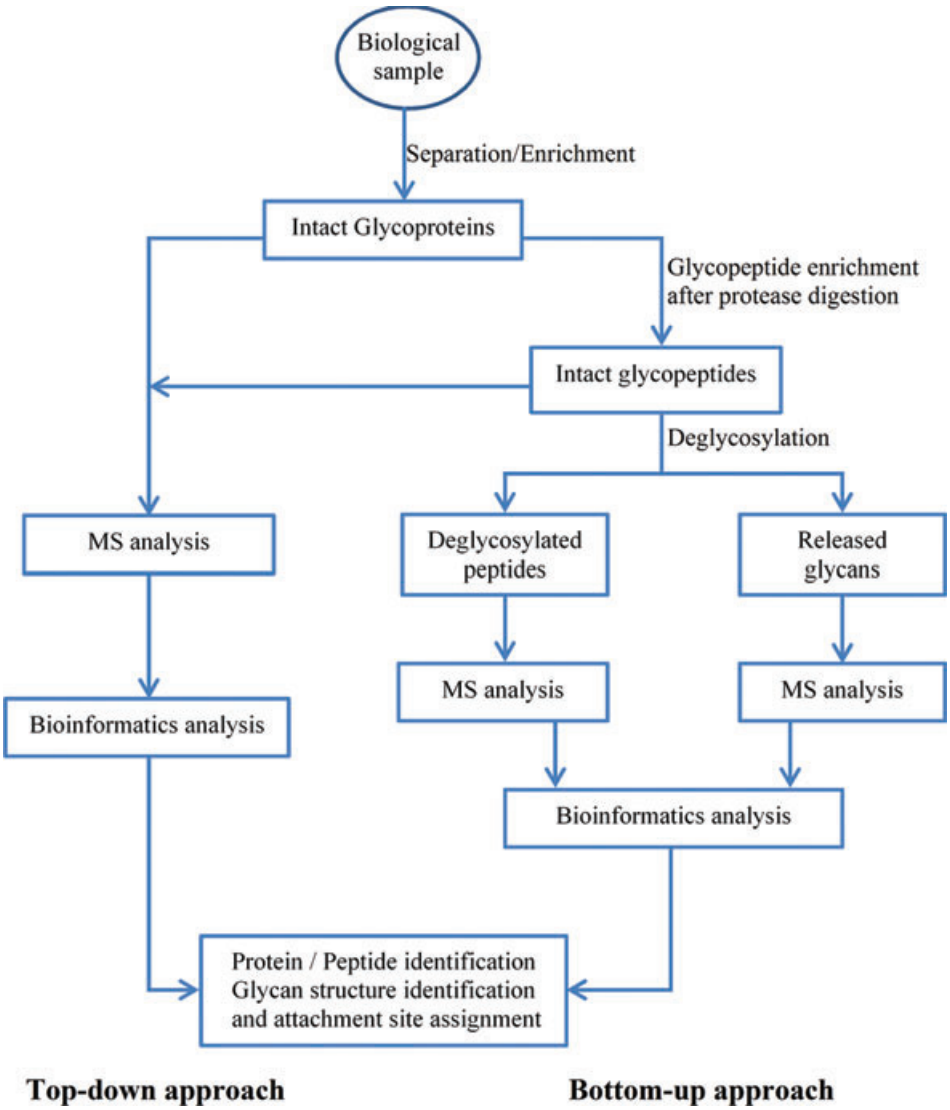

Figure 1. "Top-down" and "bottom-up" workflows in glycoproteomics.

Thus, glycosylation can be identified through the tandem spectra analysis of mass shift (the m/z increase due to the attached glycans) as glycan structures are still attached on the c'- or z' fragment ions after dissociation by ETD. For example, as shown in Fig. 2, ETD spectra have been used successfully to deduce the glycosylation sites based on the mass shift in comparison to the CID spectra from the same precursor ions [54]. Consequently, integrated CID and ETD tandem mass spectra in data-dependent LC-MS/MS are expected to play an increasingly important role in intact glycopeptide analysis [55].

In addition to MS, other approaches including HPLC, NMR, chemical reactions, radioactive labeling, as well as detection with specific lectins or antibodies have been adopted to probe the monosaccharide composition or determine the structure of glycans released from glycoproteins. As these approaches, which are often labor intensive and time consuming, have been recently reviewed elsewhere [56, 57] they will not be covered in this review.

4 Glycoinformatics: Data interpretation strategies and tools in glycoproteomics

The next important challenge in MS-based glycoproteomics is efficient interpretation of tandem mass spectra data generated from intact glycopeptide analysis using adequate bioinformatics tools. However, analyzing large amounts of data generated in high-throughput MS-based glycoproteomics experiments constitutes presently a major bottleneck in glycoproteomics research.

In general, two main strategies ("two-step" and "one-step") are most commonly utilized in glycoproteomics to ascertain glycosylation of intact glycopeptides. Two-step strategy involves stripping glycans from intact glycopeptides using enzymatic or chemical methods including widely used *N*-glycanase enzymes (e.g. PNGase F and PNGase A) to release *N*-linked glycans, or alternative chemical cleavage (e.g. β-elimination or hydrazine) to release *O*- or *N*-linked glycans

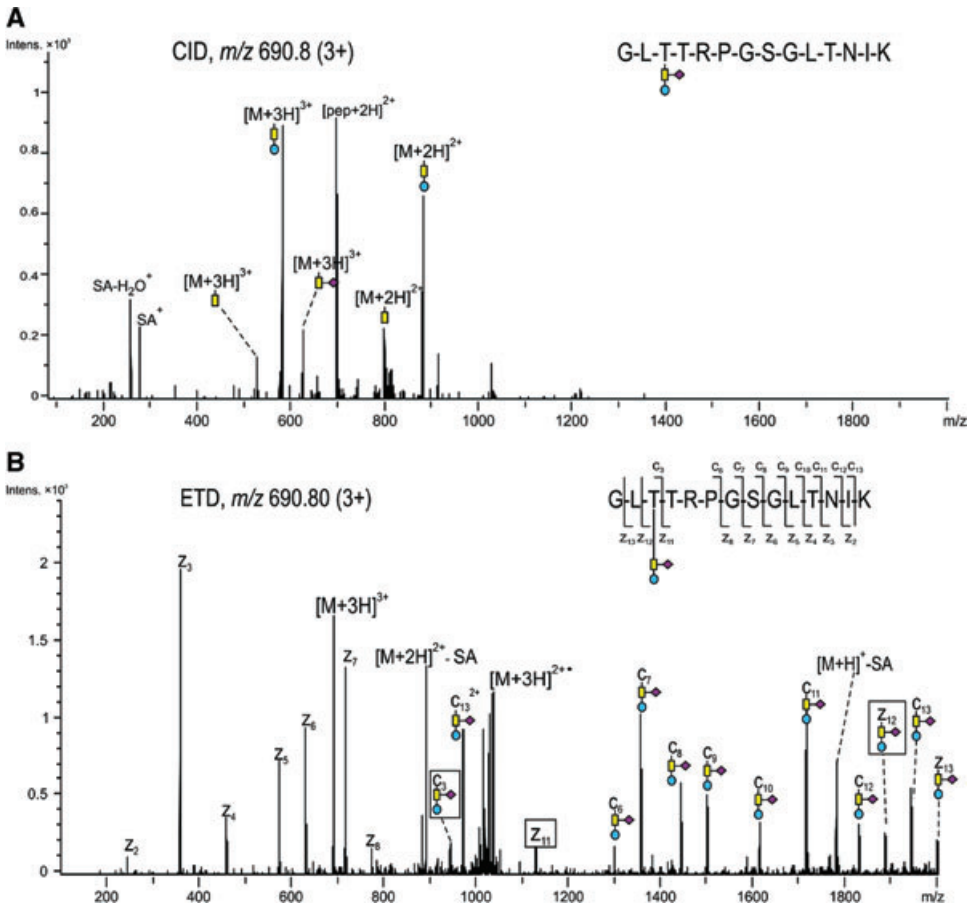

Figure 2. CID (A) and ETD (B) spectra of the precursor ion *m/z* 690.8 (3+), corresponding to glycopeptide 574–587 of the full-length APP695, showing amino acid Thr 576 occupied with the indicated Core 1 type trisaccharide. The ETD spectrum was obtained using data-dependent acquisition and the activation energy was 0.10 V. Color code: yellow, *N*-acetyl galactosamine; blue, galactose; purple, sialic acid. The fragment ions relevant for determination of the glycosylation site are indicated with black boxes. Adapted with permission from [54].

and identifying deglycosylated peptides and glycans separately by MS [58]. Releasing an attached *N*-glycan moiety by PNGase F or A results in the conversion of Asn to Asp on the attachment site, which in turn causes a mass shift for each *N*-glycosylation site on the mass spectrum. Thus, when a deglycosylated protein is further digested with trypsin, the peptides that were bound to the glycan moiety will be about 1 Da heavier than the expected theoretical mass. After subjecting these peptides to MS/MS, each peptide that possesses Asp (instead of Asn) is identified as formerly attached to the glycan moiety [59]. A similar approach involves glycan release with PNGase F in the presence of $H_2{}^{18}O$. The deglycosylated Asn will be labeled with ^{18}O and its mass altered by about 3 Da (1 Da for the Asn-to-Asp conversion and 2 Da contributed by ^{18}O) could be followed [60, 61]. However, the subsequent bioinformatics (glycoinformatics) analysis in this approach may require significant human interference with data interpretation for assigning glycan structures as information about the sites of glycan attachment cannot be inferred from MS results automatically.

The one-step strategy is to input intact glycopeptides into MS instrument and resolve peptide backbones and/or glycan structures using different dissociation modes such as CID or CID combined with ECD or ETD. In this approach, both peptide sequences could be identified and the sites and structures of glycosylation modifications assigned through the analysis of MS data using database search tools or integrated analysis software platforms. Some of the features of bioinformatics tools utilized in one-step and two-step glycoproteomics experiments to identify glycopeptide sequences and assign the sites and structures of attached glycans are summarized below.

5 Bioinformatics tools used in one-step strategy approach

IGAP (intact glycopeptide analysis pipeline) is an automated data analysis pipeline that can be utilized in one-step strategy glycoproteomics analysis for identification of *N*-glycopeptides and glycan attachment sites (Table 1). In IGAP, the raw data of intact glycopeptide MS/MS acquired without stripping down the attached glycans are used to mine for possible sites and structures of attached glycans. The raw file generated by Xcalibur is extracted to mzXML format file by ReAdw, which is a tool used for converting Thermo Scientific RAW formats into the open format mzXML and depends on Windows-only vendor libraries from Thermo.

Table 1. Partial list of software tools currently used in glycopeptide analysis utilising a one-step strategy

Software	Function	Reference[a)]	Availability
IGAP	Analysis of *N*-linked glycopeptides	[65][b)]	Free[c)]
Protein prospector	General tandem MS search engine featured with partially predefined glycosylation	http://prospector.ucsf.edu/prospector/mshome.htm	Access through the web
GlyDB (sequest)	Analysis of *N*-linked glycopeptides	[72]	NAD
Peptoonist	Analysis of *N*-linked glycopeptides with single-MS and tandem MS	[73]	NAD
N-glycopeptide library	Analysis of N-glycopeptides with custorm generated human N-glycopeptides library	[76]	NAD
GP finder (glycox)	Analysis of *N*-, or O-linked glycopeptides based on tandem MS with diagnostic ions	[74, 75]	NAD
GlycoPep ID	Analysis of *N*-linked glycosylation based on target protein and CID spectra	http://hexose.chem.ku.edu/predictiontable.php	Access through the web

a) If the software, program, or web service is an open access resource, the web address is provided, otherwise the paper reference numbers are provided.
b) This software can be downloaded through supplementary program from http://www.nature.com/nprot/journal/v6/n3/full/nprot.2010.176.html#/supplementary-information.
c) This software is free for academic use. For commercial use, the developers should be contacted.
NAD, not accessible directly, the authors of the papers should be contacted.

However, ReAdw was last released in 2009 and is no longer supported. Although, the instructions on how to install and run ReAdw could be found at the following website (http://tools.proteomecenter.org/wiki/index.php?title= Software:ReAdW), the raw format files could be also converted to the mzXML format by other format converters such as MSConvert [62]. Then the mzXML format files are used to search against database with X!Tandem open source database search engine (other search engines such as MASCOT or SEQUEST may also be used), and the identified proteins are saved as local Refinement protein database. The search results are validated with PeptideProphet and ProteinProphet based on the use of expectation-maximization algorithm to derive a mixture model of correct and incorrect peptide

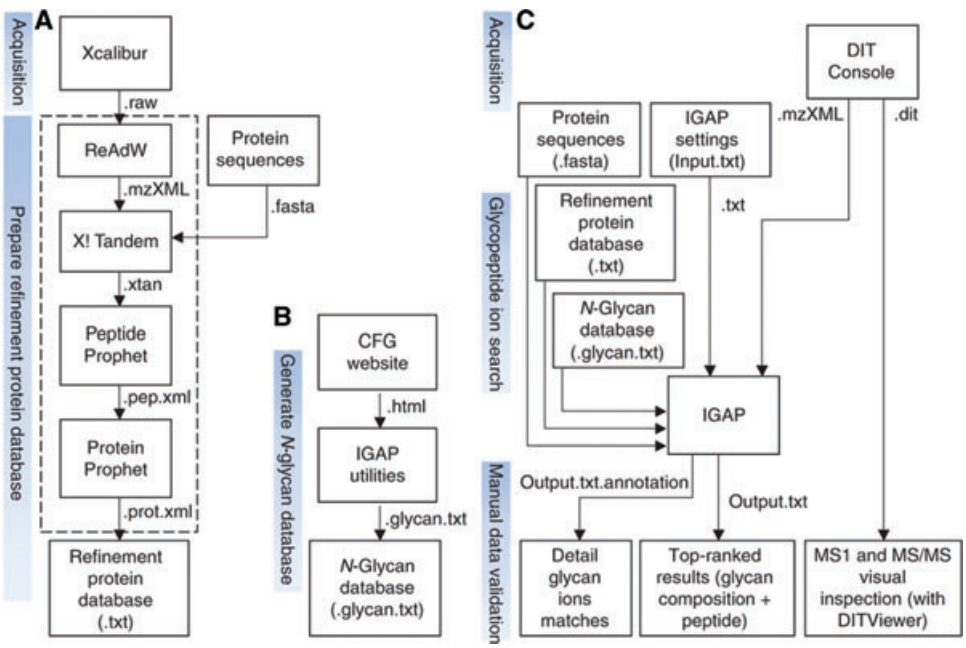

Figure 3. Bioinformatics workflow in IGAP. (A) The refinement protein database is created from the protein IDs confidently identified with the LC-ESI-MS/MS data. Each ID is recorded on a separate line. This database is experiment specific. Protein identification may be performed using other search engines. (B) The *N*-glycan database is constructed with the data from consortium for functional glycomics (CFG) glycan structures database. This database only needs to be updated when new entries are found at CFG. (C) MALDI-DIT MS/MS spectrum is searched by IGAP, which generates the top 20 glycan composition results in the tab-delimited file "Output.txt" and the matched peaks annotation in "Output.txt.annotation." Manual validation of the results is performed. DITViewer provides visual access to the acquired spectra. Reproduced with permission from [65].

identifications from the data [63]. Next, IGAP utilities are used to construct local *N*-Glycan database based on CFG (consortium for functional glycomics) glycan structure database, which offers detailed structural and chemical information on thousands of glycans, including both synthetic glycans and glycans isolated from biological sources [64]. For example, as of May 2012, searching the glycan structures database with the *N*-linked "core" substructure containing three mannose and two *N*-acetylglucosamine residues returns 4605 *N*-linked glycan records [http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/structure/searchThisStructure.jsp?lincode=Ma3%28Ma6%29Mb4GNb4GN]. The acquired glycopeptide MS/MS spectra are processed through IGAP to compute the possible theoretical glycopeptides based on protein sequences from local Refinement protein database and local CFG glycan database. IGAP in silico digests the theoretical tryptic peptides from the Refinement protein database with 0, 1 and 2 missed cleavages, and only peptides containing a consensus sequence N-X-S/T (X except proline) are considered as potential *N*-glycosylated peptides. IGAP considers the fragmentation of the glycosidic bonds (with Y ions and X^{0-2} of *N*-glycans on the intact potential *N*-linked glycopeptides) and the fragmentation of the peptide bonds (with both b and y series ions of the potential *N*-linked glycopeptides along with an attached GlcNAc). Then these two kinds of fragment ions are merged together to generate the theoretical MS/MS spectrum. Experimental spectrum is cumulative intensity normalized, the measured *m/z* range is divided into 100 Th (Thomson) regions, the eight most intensive peaks in each Th region is extracted to form the intensity spectrum, which is used to match against the theoretical spectrum. In this process, A-Score algorithm is used to score the matching, and the collected matches are ranked by glycan moiety score, glycopeptide probability of random matches, and glycopeptide score. The top 20 glycan composition results are retained and further narrowed down based on other available information. Thus, without striping down the attached glycans, data generated by MS/MS could be used to identify intact glycopeptide sequences and the sites and structures of the attached glycans based on in-depth tandem mass spectra data mining technologies. Typical bioinformatics workflow in IGAP is depicted in Fig. 3. IGAP program can be obtained from the Nature website (http://www.nature.com/nprot/journal/v6/n3/full/nprot.2010.176.html#supplementary-information) from the supplementary files of the following article [65]. However, currently IGAP can be used for *N*-glycoproteomics analysis only.

Protein prospector is also a tandem mass spectra database search software used for identifying the glycosylation sites in one-step strategy glycoproteomics analysis. It has been optimized and successfully utilized for *O*-glycoproteomics [66, 67]. MS data acquired using CID combined with ECD or ETD have been used to identify O-GlcNAc modification sites on native peptides with concomitant identification of peptides sequences by the protein prospector database search engine [5, 68, 69]. Compared with ECD, ETD markedly increases the number of O-GlcNAc modification sites determined in a single experiment, which could be related to a higher charge and lower *m/z* components in the ETD fragmentation process [70]. Unlike many of the modification site scoring tools implemented in other database search engines, the Batch-Tag scoring used in protein prospector is applicable to all modifications, and the scoring algorithm named SLIP (site localization in peptide) has been designed to calculate and compare probability and expectation values for the same peptide with different site assignments [71].

GlyDB annotates tandem mass spectra of *N*-linked glycopeptides using an in-house custom-built linearized glycan structure database and utilizes a general peptide database search engine Sequest to assign experimental tandem mass spectra to individual glycoforms [72]. Some other in-house developed software tools or scripts like Peptoonist, GlycoPep ID, GlycoX, or the upgrade version of GP finder can also be used for interpreting glycopeptide MS data in one-step strategy glycoproteomics [73–77].

In addition to using fully automated software tools like IGAP or Protein Prospector, identification of many *N*-, and *O*-glycosylation sites was reported based on manual analysis of tandem mass spectra or using in-house developed scripts. For example, in the analysis of *O*-glycosylation, diagnostic glycan oxoniums such as *m/z* 163 (Hex+), *m/z* 292 (NeuAc+), *m/z* 204 (HexNAc+), or *m/z* 366 (HexHexNAc+) will be formed in the tandem mass spectra generated from B-type and Y-type cleavages of glycosidic bonds [78, 79]. By tracking these diagnostic glycan oxoniums in the MS2, the glycosidic linkage fragmentation patterns of MS2 and MS3 spectra can be used to analyze and determine the sites and structures of the attached glycans. However, this kind of manual interpretation is time consuming and potential errors including false positives and false negatives cannot be accounted for due to the lack of statistical validation process.

6 Bioinformatics tools used in two-step strategy approach

Two-step strategy is another glycosylation analysis strategy commonly used in glycoproteomics research. In this approach, glycans are stripped down from intact glycopeptides and then "deglycosylated" peptides and recovered glycans identified separately by MS. In this approach, glycopeptide backbone sequences and the structures of the attached glycans could be identified, but information regarding the sites of glycan attachment cannot be acquired [80]. The first step or the process of identifying "deglycosylated" peptide sequences is the same that is used in general proteomics research, and the software tools or algorithms for analyzing the MS/MS data have been well reviewed [81–84]. Mascot and Sequest are currently the two database search engines most commonly used for identifying protein sequences with MS, which are

Table 2. General database search engines used in MS/MS

Software	Function	Reference	Availability
MASCOT	MS data search engine	http://www.matrixscience.com/	Commercial
OMSSA	MS data search engine	http://pubchem.ncbi.nlm.nih.gov/omssa/	Open source free
Protein prospector	MS data search engine	http://prospector.ucsf.edu/prospector/mshome.htm	Access through the web
SEQUEST	MS data search engine	http://fields.scripps.edu/sequest/	Commercial
Spectrum mill	MS data search engine	http://www.chem.agilent.com/	Commercial
X!Tandem	MS data search engine	http://www.thegpm.org/tandem/	Open source free

publicly available (Table 2). Other open source software tools like X!tandem or OMSSA are also available and could be downloaded and installed on local computers offering much flexibility in configuring the search parameters and databases.

The second step in the two-step strategy glycopeptide analysis is to identify the structures of the separated glycans. Although automatic interpretation of glycan MS data is still a challenging task, there are several software applications available to deduce probable carbohydrate compositions from MS data [85, 86] generated from a single round mass analysis or a tandem mass analysis (MS/MS or higher order (MS^n)). Consequently, these glycoinformatics tools can be generally divided into single-MS glycan analysis software and/or tandem MS glycan analysis software (Table 3).

GlycoMod is one of the first such tools developed to compute all possible glycan compositions from experimentally derived mass spectra data by comparing the actual mass of the glycan to a list of precomputed masses of glycan compositions thus allowing for the composition of a glycan attached to a peptide to be computed if the sequence or the mass of the peptide is known [87]. Cartoonist is another program designed for computing glycans based on data generated from a single round of mass analysis, which is used by CFG to profile glycans from various organisms and tissues [88, 89]. Glypeps is the software that allows unraveling information encrypted in the deltamass value of accurate peptide masses. When the deltamass value is indicative of a glycopeptide, Glypeps could be used to compute a list of proposed *N*-glycan structures if the sequence of the peptide is known [90]. GlycoSpectrumScan is a web-based tool to identify the glycoheterogeneity on a peptide from mass spectra data. It uses single-MS data and two experimental datasets, including oligosaccharide compositions of the *N*- and/or *O*-linked glycans present in the sample and in silico derived peptide masses of proteolytically digested proteins, to identify glycopeptides and determine the relative distribution of *N*- and *O*-glycoforms at each site [91].

The structures of the attached glycans could be successfully identified with MS/MS as well. This also requires the use of advanced data interpretation tools to decipher complicated glycan structures. GlycoFragment and GlycoSearchMS are two web tools available to compute all theoretically possible fragments of complex carbohydrates based on MS/MS data. GlycoFragment computes all theoretically possible MS-relevant fragments of oligosaccharides as defined by the extended IUPAC nomenclature, while GlycoSearchMS takes the experimental mass spectra peak values as an input and searches for matches with the calculated fragments of all structures contained in the SweetDB database [92]. GlycosidIQ interprets oligosaccharide mass spectra based on matching experimental data with theoretically fragmented oligosaccharides generated from the database GlycoSuiteDB [93]. GlycoWorkbench is a tool developed by the EUROCarbDB initiative, which is designed to provide support for the routine interpretation of MS data. It evaluates a set of structures proposed by the user via matching the list of peaks derived from the tandem mass spectra against the corresponding theoretical list of fragment masses [7, 94]. Glyquest determines asparagine-linked glycan (*N*-glycan) structures based on tandem mass spectra of glycopeptides using a built-in *N*-glycan structure database and an integrated database search engine [95]. Glycominer identifies *N*-glycans from tandem mass spectra based on an empirical algorithm, which determines the low mass oxonium ions, deduces oligosaccharide losses from the protonated molecule, and identifies the mass of the peptide residue [96]. SimGlycan™ is a desktop tool designed to predict the glycan structures from MS/MS spectra through using database searching and propriety scoring algorithm against its own database of theoretical fragmentation of over 9000 glycans. SimGlycan™ can predict the attached glycan structures if the mass of the peptide or the peptide sequence is known [97].

In contrast to adopting a strategy of matching the oligosaccharide mass spectra with databases of the theoretically fragmented ones employed in bioinformatics tools mentioned above, several software tools interpret tandem mass spectra data based on de novo glycan sequencing. STAT is a web-based tool for saccharide topology analysis (Table 4). It extracts information from a set of MS^n spectra and computes all possible structures, which are generated and evaluated against the MS^n data so the list of possible structures is assigned a rating based on the likelihood that it is the correct sequence [98]. GlySpy is the prototype tool that implements the OSCAR algorithms. It accepts user-selected MS^n ion fragment paths and applies logical constraints to produce the full set of glycan structures that could yield the selected ions [99]. StrOligo algorithm first builds a relationship tree accounting for each observed loss of a monosaccharide moiety and then evaluates the agreement between the tree and each proposed possible structure from combinations of adducts and fragment ion types generated by MS/MS with a score.

Table 3. Partial list of software tools currently used in glycopeptide analysis utilising a two-step strategy

Software	Function	Reference[a)]	Availability
Tools for assigning glycan structures based on single-MS			
GlycoMod	Predict glycan strutures based on single-MS	http://web.expasy.org/glycomod	Access through the web
Cartoonist	Annotate permethylated *N*-glycans with single-MS	[88, 89]	NAD
GlyPeps	Annotate *N*-glycans with single-MS when peptide sequences are known	http://www.glycosciences.de/spec/glypeps/	Access through the web
GlycoSpectrumScan	Web-based tool to identify the glycoheterogeneity on a peptide from MS data	http://www.glycospectrumscan.org/	Access through the web
Tools for assigning glycan structures based on tandem MS			
GlycoFragment	Annotate glycan structures based on theoretically possible fragments and tandem MS	http://www.glycosciences.de/tools/GlycoFragments/	Access through the web
GlycoSearchMS	Annotate glycan structures based on theoretically possible fragments from SweetDB	http://www.glycosciences.de/sweetdb/start.php?action=form_ms_search	Access through the web
GlycosidIQ	Annotate glycan structures based on theoretically possible fragments from GlycoSuite DB	http://glycosuitedb.expasy.org/glycosuite/query	Access through the web
GlycoWorkbench	Design for rapid drawing of glycan structures and it can automatically match to tandem MS data	http://www.glycoworkbench.org/	Free
GlyQuest	Determine *N*-glycan structures with built-in database and search engine	[95]	NAD
GlycoMiner	Determine *N*-glycans from tandem MS data with empirical algorithm	http://www.chemres.hu/ms/glycominer/index.php	Free

a) If the software, program, or web service is an open access resource, the web address is provided, otherwise the paper reference numbers are provided.
NAD, not accessible directly, the authors of the papers should be contacted.

Subsequently, the best combination is selected based on the score and the relevant peaks are labeled in the experimental mass spectrum using a modified nomenclature [100]. GlycoMaster uses heuristic dynamic programing technique to compute the best possible sequence structure among all possible monosaccharide combinations [101]. GLYCH interprets tandem mass spectra of oligosaccharides based on the appearance pattern of cross-ring ions taking into account double fragment ions as well [102]. Glyco-Peakfinder is a tool for a fast annotation of glycan MS spectra, which provides the option of detecting differently and/or multiply charged ions in one calculation cycle accounting as well for modifications in the reducing ends or within the sequences of oligosaccharides [103].

7 Other issues related to glycan analysis pipeline and concluding remarks

The interpretation and implementation of bioinformatic solutions for glycopeptide data generated by MS/MS is still a very challenging task due to the overlap in both peptide and glycan fragmentation. The evaluation of the matches between experimental and theoretical spectra (peptide-spectrum matches, PSMs) is vital for correctly identifying the PSMs with statistical confidence, especially in glycoproteomics based on the bottom-up pipeline. However, to the best of our knowledge, to date only two algorithms have been implemented in the evaluation of PSMs in glycoproteomics. IGAP evaluates the matching between experimental spectra against combined theoretical spectra with the A-score algorithm, originated from phosphorylation analysis, to discriminate the *N*-linked glycopeptides from unmodified peptides. Similarly, OScore algorithm has been reported to validate the identification of O-GlcNAc modified peptides generated from data-dependent ETD tandem experiments [104]. However, due to the extreme complexity of glycosylation, additional efforts are urgently needed to develop more reliable tools and algorithms assisting with the evaluation of PSMs, assignment of the glycosylation sites on glycopeptides, and identification of the structures of the attached glycan moieties.

As glycan structures cannot be predicted from theoretical sequence databases, no matter which glycan site assignment strategy is used, a glycomics database is necessary for glycan identification in MS-based glycoproteomics. Currently, there are several carbohydrate databases developed and maintained by different academic and commercial organizations including CFG Glycan database, GLYCOSCIENCES.de Glycan database, KEGG Glycan database, and other glycan databases and resources such as EUROCarbDB, UniCarb-DB, GMDB, glycosuite database, glycominds, bacterial carbohydrate

Table 4. Partial list of software tools currently used for de novo sequencing based on tandem MS

Software	Function	Reference[a)]	Availability
STAT	De novo assign tandem MS spectra for an oligosaccharide of up to 10 residues	[98]	NAD
GlySpy (OSCAR)	Assign glycan structures based on user-selected MS^n ion fragment paths logical constraints	[99]	NAD
StrOligo	Interpret tandem MS data with relationship tree model to fit the experimental data	[100]	NAD
GlycoMaster	De novo assign N-glycan structures with heuristic dynamic programming technique and Branch-and-Bound algorithm	[101]	NAD
GLYCH	Interpret glycan strucutres from tandem MS data based on the appearance pattern of cross-ring ions and dynamic programming algorithm	[102]	NAD
[b)]Glyco-Peakfinder	De novo assign tandem MS data with all types of fragment ions including monosaccharide cross-ring cleavage products and multiply charged ions	http://www.glyco-peakfinder.org/	Access through the web

a) If the software, program, or web service is an open access resource, the web address is provided, otherwise the paper reference numbers are provided.
b) This tool can be accessed through the website http://www.glyco-peakfinder.org/.
NAD, not accessible directly, the authors of the papers should be contacted.

structural databases (BCSCD), CAZy (Carbohydrate-Active enZYmes Database), which have been well introduced and/or reviewed elsewhere [7, 23, 105, 106]. However, due to the different data formats used to encode carbohydrate structures, there is almost no direct cross-referencing between these established carbohydrate databases leading to the existence of multiple disconnected and incompatible islands in glycomics. Although, several efforts to correct this situation (generation of GlyDE data exchange standard, GlycoCT sequence format, GlycomeDB database) have been reported [85, 107–109], currently it presents major roadblocks for efficient communication and data sharing within the glycoscience community. In this situation, a closer collaboration regarding the development of glycobioinfomatic concepts between major North-American, European, and Asian bioinformatics centers could be expected to eliminate the lack of commonly recognized standards for glycan definition formats, data exchange formats, and data share databases. The 2nd Beilstein Symposium on glycobioinformatics "Cracking the Sugar Code by Navigating the Glycospace" showed trends of cooperation between several big bioinformatics centers and toward integration of glycoprotein resources (http://www.beilstein-institut.de/en/symposia/overview/proceedings/2011-glycobioinformatics/). The new initiative UniCarbKB represents recent efforts of glycobioinformatics community toward integrating different resources into one universal glycomics knowledgebase, which could provide a comprehensive publically accessible catalogue of information about carbohydrates [110]. Also, developing proteomic resources like Tranche repository should improve the data exchange and tool share in glycoproteomics as well [111]. EUROcarbDB project has published their software libraries and bioinformatics tools through googlecode at http://eurocarb.googlecode.com, under the terms of the Lesser General Public License like many other open source projects. However, many glycobioinformatics tools still can only be downloaded from the developer's websites or through the paper's attachments. Enhancing rapidly developing field of glycobioinformatics, which provides vital support for glycomics and glycoproteomics research, with new comprehensive and universal tools for data mining and structural analysis will greatly improve glycopeptide decoding.

This work was supported in parts by Award Number 1I01BX000609 from the Biomedical Laboratory Research & Development Service of the VA Office of Research and Development to VVG and American Heart Association National SDG 0830287N to OVG.

The authors have declared no conflict of interest.

8 References

[1] Jensen, O. N., Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.* 2006, *7*, 391–403.

[2] Walsh, C. T., *Posttranslational Modification of Proteins: Expanding Nature's Inventory*, Roberts and Company Publishers, Greenwood Village 2006.

[3] Ribet, D., Cossart, P., Pathogen-mediated posttranslational modifications: a re-emerging field. *Cell* 2010, *143*, 694–702.

[4] Walsh, C. T., Garneau-Tsodikova, S., Gatto, G. J., Jr., Protein posttranslational modifications: the chemistry of

proteome diversifications. *Angew. Chem. Int. Ed. Engl.* 2005, *44*, 7342–7372.

[5] Chalkley, R. J., Thalhammer, A., Schoepfer, R., Burlingame, A. L., Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 8894–8899.

[6] Turnbull, J. E., Field, R. A., Emerging glycomics technologies. *Nat. Chem. Biol.* 2007, *3*, 74–77.

[7] von der Lieth, C. W., Freire, A. A., Blank, D., Campbell, M. P. et al., EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology* 2011, *21*, 493–502.

[8] Hart, G. W., Copeland, R. J., Glycomics hits the big time. *Cell* 2010, *143*, 672–676.

[9] Schauer, R., Sialic acids as regulators of molecular and cellular interactions. *Curr. Opin. Struct. Biol.* 2009, *19*, 507–514.

[10] Spiro, R. G., Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* 2002, *12*, 43R–56R.

[11] Schachter, H., Freeze, H. H., Glycosylation diseases: quo vadis? *Biochim. Biophys. Acta* 2009, *1792*, 925–930.

[12] Hansen, S. F., Bettler, E., Rinnan, A., Engelsen, S. B. et al., Exploring genomes for glycosyltransferases. *Mol. Biosyst.* 2010, *6*, 1773–1781.

[13] Narimatsu, H., Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj. J.* 2004, *21*, 17–24.

[14] Khoury, G. A., Baliban, R. C., Floudas, C. A., Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.* 2011, *1*, 90.

[15] 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.

[16] Colgrave, M. L., Snelling, H. J., Shiell, B. J., Feng, Y. R. et al., Site occupancy and glycan compositional analysis of two soluble recombinant forms of the attachment glycoprotein of Hendra virus. *Glycobiology* 2012, *22*, 572–584.

[17] Arnold, J. N., Wormald, M. R., Sim, R. B., Rudd, P. M., Dwek, R. A., The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu. Rev. Immunol.* 2007, *25*, 21–50.

[18] Zielinska, D. F., Gnad, F., Wisniewski, J. R., Mann, M., Precision mapping of an in vivo *N*-glycoproteome reveals rigid topological and sequence constraints. *Cell* 2010, *141*, 897–907.

[19] Bertozzi, C. R., Sasisekharan, R., in: Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E. (Eds.), *Essentials of Glycobiology*, 2nd Edn., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY 2009, pp. 679–690.

[20] Varki, A., Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. *Nature* 2007, *446*, 1023–1029.

[21] Tian, Y., Zhang, H., Glycoproteomics and clinical applications. *Proteomics Clin. Appl.* 2010, *4*, 124–132.

[22] Rademacher, T. W., Parekh, R. B., Dwek, R. A., *Glycobiology. Annu. Rev. Biochem.* 1988, *57*, 785–838.

[23] Raman, R., Raguram, S., Venkataraman, G., Paulson, J. C. et al., Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat. Methods* 2005, *2*, 817–824.

[24] Moremen, K. W., Molinari, M., *N*-linked glycan recognition and processing: the molecular basis of endoplasmic reticulum quality control. *Curr. Opin. Struct. Biol.* 2006, *16*, 592–599.

[25] Maattanen, P., Gehring, K., Bergeron, J. J., Thomas, D. Y., Protein quality control in the ER: the recognition of misfolded proteins. *Semin. Cell Dev. Biol.* 2010, *21*, 500–511.

[26] Mulloy, B., Hart, G. W., Stanley, P., in: Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E. (Eds.), *Essentials of Glycobiology*, 2nd Edn., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY 2009, pp. 661–678.

[27] Dempski, R. E., Jr., Imperiali, B., Oligosaccharyl transferase: gatekeeper to the secretory pathway. *Curr. Opin. Chem. Biol.* 2002, *6*, 844–850.

[28] Stanley, P., Schachter, H., Taniguchi, N., in: Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E. (Eds.), *Essentials of Glycobiology*, 2nd Edn., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY 2009, pp. 101–114.

[29] McDonald, C. A., Yang, J. Y., Marathe, V., Yen, T. Y. et al., Combining results from lectin affinity chromatography and glycocapture approaches substantially improves the coverage of the glycoproteome. *Mol. Cell Proteomics* 2009, *8*, 287–301.

[30] Brockhausen, I., Schachter, H., Stanley, P., in: Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E. (Eds.), *Essentials of Glycobiology*, 2nd Edn., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY 2009, pp. 115–128.

[31] Hart, G. W., Slawson, C., Ramirez-Correa, G., Lagerlof, O., Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. *Annu. Rev. Biochem.* 2011, *80*, 825–858.

[32] Hanover, J. A., Krause, M. W., Love, D. C., Bittersweet memories: linking metabolism to epigenetics through O-GlcNAcylation. *Nat. Rev. Mol. Cell Biol.* 2012, *13*, 312–321.

[33] Hart, G. W., Housley, M. P., Slawson, C., Cycling of O-linked beta-*N*-acetylglucosamine on nucleocytoplasmic proteins. *Nature* 2007, *446*, 1017–1022.

[34] Zhang, Q., Ames, J. M., Smith, R. D., Baynes, J. W. et al., A perspective on the Maillard reaction and the analysis of protein glycation by mass spectrometry: probing the pathogenesis of chronic disease. *J. Proteome Res.* 2009, *8*, 754–769.

[35] Pan, S., Chen, R., Aebersold, R., Brentnall, T. A., Mass spectrometry based glycoproteomics–from a proteomics perspective. *Mol. Cell Proteomics* 2011, *10*, R110 003251.

[36] Lazar, I. M., Lazar, A. C., Cortes, D. F., Kabulski, J. L., Recent advances in the MS analysis of glycoproteins: theoretical considerations. *Electrophoresis* 2011, *32*, 3–13.

[37] Chait, B. T., Chemistry. Mass spectrometry: bottom-up or top-down? *Science* 2006, *314*, 65–66.

[38] Borchers, C. H., Thapar, R., Petrotchenko, E. V., Torres, M. P. et al., Combined top-down and bottom-up proteomics identifies a phosphorylation site in stem-loop-binding proteins that contributes to high-affinity RNA binding. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 3094–3099.

[39] Wu, S., Tolic, N., Tian, Z., Robinson, E. W. et al., An integrated top-down and bottom-up strategy for characterization of protein isoforms and modifications. *Methods Mol. Biol.* 2011, *694*, 291–304.

[40] Carpentieri, A., Giangrande, C., Pucci, P., Amoresano, A., Glycoproteome study in myocardial lesions serum by integrated mass spectrometry approach: preliminary insights. *Eur. J. Mass Spectrom. (Chichester, Eng)* 2010, *16*, 123–149.

[41] Chen, G., Pramanik, B. N., Application of LC/MS to proteomics studies: current status and future prospects. *Drug Discov. Today* 2009, *14*, 465–471.

[42] Siuti, N., Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* 2007, *4*, 817–821.

[43] Palumbo, A. M., Smith, S. A., Kalcic, C. L., Dantus, M. et al., Tandem mass spectrometry strategies for phosphoproteome analysis. *Mass Spectrom. Rev.* 2011, *30*, 600–625.

[44] Wuhrer, M., Catalina, M. I., Deelder, A. M., Hokke, C. H., Glycoproteomics based on tandem mass spectrometry of glycopeptides. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2007, *849*, 115–128.

[45] Greis, K. D., Hayes, B. K., Comer, F. I., Kirk, M. 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.

[46] Ito, H., Takegawa, Y., Deguchi, K., Nagai, S. et al., Direct structural assignment of neutral and sialylated *N*-glycans of glycopeptides using collision-induced dissociation MSn spectral matching. *Rapid Commun. Mass Spectrom.* 2006, *20*, 3557–3565.

[47] Renfrow, M. B., Mackay, C. L., Chalmers, M. J., Julian, B. A. et al., Analysis of O-glycan heterogeneity in IgA1 myeloma proteins by Fourier transform ion cyclotron resonance mass spectrometry: implications for IgA nephropathy. *Anal. Bioanal. Chem.* 2007, *389*, 1397–1407.

[48] Kjeldsen, F., Haselmann, K. F., Budnik, B. A., Sorensen, E. S. et al., Complete characterization of posttranslational modification sites in the bovine milk protein PP3 by tandem mass spectrometry with electron capture dissociation as the last stage. *Anal. Chem.* 2003, *75*, 2355–2361.

[49] Zubarev, R. A., Kelleher, N. L., McLafferty, F. W., Electron capture dissociation of multiply charged protein cations. A non-ergodic process. *J. Am Chem. Soc.* 1998, *120*, 3265–3266.

[50] Mirgorodskaya, E., Roepstorff, P., Zubarev, R. A., Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal. Chem.* 1999, *71*, 4431–4436.

[51] Baba, T., Hashimoto, Y., Hasegawa, H., Hirabayashi, A. et al., Electron capture dissociation in a radio frequency ion trap. *Anal. Chem.* 2004, *76*, 4263–4266.

[52] Mikesh, L. M., Ueberheide, B., Chi, A., Coon, J. J. et al., The utility of ETD mass spectrometry in proteomic analysis. *Biochim. Biophys. Acta* 2006, *1764*, 1811–1822.

[53] Wiesner, J., Premsler, T., Sickmann, A., Application of electron transfer dissociation (ETD) for the analysis of post-translational modifications. *Proteomics* 2008, *8*, 4466–4483.

[54] Perdivara, I., Petrovich, R., Allinquant, B., Deterding, L. J. et al., Elucidation of O-glycosylation structures of the beta-amyloid precursor protein by liquid chromatography-mass spectrometry using electron transfer dissociation and collision induced dissociation. *J. Proteome Res.* 2009, *8*, 631–642.

[55] Harvey, D. J., Proteomic analysis of glycosylation: structural determination of N- and O-linked glycans by mass spectrometry. *Expert. Rev. Proteomics* 2005, *2*, 87–101.

[56] Geyer, H., Geyer, R., Strategies for analysis of glycoprotein glycosylation. *Biochim. Biophys. Acta* 2006, *1764*, 1853–1869.

[57] Rakus, J. F., Mahal, L. K., New technologies for glycomic analysis: toward a systematic understanding of the glycome. *Annu. Rev. Anal. Chem. (Palo Alto Calif)* 2011, *4*, 367–392.

[58] Goetz, J. A., Novotny, M. V., Mechref, Y., Enzymatic/chemical release of O-glycans allowing MS analysis at high sensitivity. *Anal. Chem.* 2009, *81*, 9546–9552.

[59] Roth, Z., Parnes, S., Wiel, S., Sagi, A. et al., N-glycan moieties of the crustacean egg yolk protein and their glycosylation sites. *Glycoconj. J.* 2010, *27*, 159–169.

[60] Kuster, B., Mann, M., 18O-labeling of N-glycosylation sites to improve the identification of gel-separated glycoproteins using peptide mass mapping and database searching. *Anal. Chem.* 1999, *71*, 1431–1440.

[61] Kaji, H., Saito, H., Yamauchi, Y., Shinkawa, T. et al., Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat. Biotechnol.* 2003, *21*, 667–672.

[62] Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T. et al., A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 2010, *10*, 1150–1159.

[63] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646–4658.

[64] Raman, R., Venkataraman, M., Ramakrishnan, S., Lang, W. et al., Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology* 2006, *16*, 82R–90R.

[65] Wang, H., Wong, C. H., Chin, A., Taguchi, A. et al., Integrated mass spectrometry-based analysis of plasma glycoproteins and their glycan modifications. *Nat. Protoc.* 2011, *6*, 253–269.

[66] Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C. et al., Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in

Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol. Cell Proteomics* 2005, *4*, 1194–1204.

[67] Chalkley, R. J., Baker, P. R., Medzihradszky, K. F., Lynn, A. J. et al., In-depth analysis of tandem mass spectrometry data from disparate instrument types. *Mol. Cell Proteomics* 2008, *7*, 2386–2398.

[68] Darula, Z., Chalkley, R. J., Baker, P., Burlingame, A. L. et al., Mass spectrometric analysis, automated identification and complete annotation of O-linked glycopeptides. *Eur. J. Mass Spectrom. (Chichester, Eng)* 2010, *16*, 421–428.

[69] Vosseller, K., Trinidad, J. C., Chalkley, R. J., Specht, C. G. et al., O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. *Mol. Cell Proteomics* 2006, *5*, 923–934.

[70] Baker, P. R., Medzihradszky, K. F., Chalkley, R. J., Improving software performance for peptide electron transfer dissociation data analysis by implementation of charge state- and sequence-dependent scoring. *Mol. Cell Proteomics* 2010, *9*, 1795–1803.

[71] Baker, P. R., Trinidad, J. C., Chalkley, R. J., Modification site localization scoring integrated into a search engine. *Mol. Cell Proteomics* 2011, *10*, M111 008078.

[72] Ren, J. M., Rejtar, T., Li, L., Karger, B. L., *N*-Glycan structure annotation of glycopeptides using a linearized glycan structure database (GlyDB). *J. Proteome Res.* 2007, *6*, 3162–3173.

[73] Goldberg, D., Bern, M., Parry, S., Sutton-Smith, M. et al., Automated *N*-glycopeptide identification using a combination of single- and tandem-MS. *J. Proteome Res.* 2007, *6*, 3995–4005.

[74] Nwosu, C. C., Seipert, R. R., Strum, J. S., Hua, S. S. et al., Simultaneous and extensive site-specific N- and O-glycosylation analysis in protein mixtures. *J. Proteome Res.* 2011, *10*, 2612–2624.

[75] An, H. J., Tillinghast, J. S., Woodruff, D. L., Rocke, D. M. et al., A new computer program (GlycoX) to determine simultaneously the glycosylation sites and oligosaccharide heterogeneity of glycoproteins. *J. Proteome Res.* 2006, *5*, 2800–2808.

[76] Joenvaara, S., Ritamo, I., Peltoniemi, H., Renkonen, R., N-glycoproteomics—an automated workflow approach. *Glycobiology* 2008, *18*, 339–349.

[77] Irungu, J., Go, E. P., Dalpathado, D. S., Desaire, H., Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID. *Anal. Chem.* 2007, *79*, 3065–3074.

[78] Nilsson, J., Larson, G., Grahn, A., Characterization of site-specific O-glycan structures within the mucin-like domain of alpha-dystroglycan from human skeletal muscle. *Glycobiology* 2010, *20*, 1160–1169.

[79] Chen, Y., Liu, M., Yan, G., Lu, H. et al., One-pipeline approach achieving glycoprotein identification and obtaining intact glycopeptide information by tandem mass spectrometry. *Mol. Biosyst.* 2010, *6*, 2417–2422.

[80] Wada, Y., Dell, A., Haslam, S. M., Tissot, B. et al., Comparison of methods for profiling O-glycosylation: human proteome organisation human disease glycomics/proteome initiative multi-institutional study of IgA1. *Mol. Cell Proteomics* 2010, *9*, 719–727.

[81] Kapp, E., Schutz, F., Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Curr. Protoc. Protein Sci.* 2007, *49*, 25.2.1–25.2.19.

[82] Nesvizhskii, A. I., Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol. Biol.* 2007, *367*, 87–119.

[83] Lu, B., Xu, T., Park, S. K., Yates, J. R., 3rd, Shotgun protein identification and quantification by mass spectrometry. *Methods Mol. Biol.* 2009, *564*, 261–288.

[84] Nesvizhskii, A. I., Vitek, O., Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 2007, *4*, 787–797.

[85] Frank, M., Schloissnig, S., Bioinformatics and molecular modeling in glycobiology. *Cell Mol. Life Sci.* 2010, *67*, 2749–2772.

[86] Perez, S., Mulloy, B., Prospects for glycoinformatics. *Curr. Opin. Struct. Biol.* 2005, *15*, 517–524.

[87] Cooper, C. A., Gasteiger, E., Packer, N. H., GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* 2001, *1*, 340–349.

[88] Goldberg, D., Sutton-Smith, M., Paulson, J., Dell, A., Automatic annotation of matrix-assisted laser desorption/ionization *N*-glycan spectra. *Proteomics* 2005, *5*, 865–875.

[89] Goldberg, D., Bern, M., North, S. J., Haslam, S. M., Dell, A., Glycan family analysis for deducing *N*-glycan topology from single MS. *Bioinformatics* 2009, *25*, 365–371.

[90] Lehmann, W. D., Bohne, A., von Der Lieth, C. W., The information encrypted in accurate peptide masses-improved protein identification and assistance in glycopeptide identification and characterization. *J. Mass Spectrom.* 2000, *35*, 1335–1341.

[91] Deshpande, N., Jensen, P. H., Packer, N. H., Kolarich, D., GlycoSpectrumScan: fishing glycopeptides from MS spectra of protease digests of human colostrum sIgA. *J. Proteome Res.* 2010, *9*, 1063–1075.

[92] Lohmann, K. K., von der Lieth, C. W., GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res.* 2004, *32*, W261–266.

[93] Joshi, H. J., Harrison, M. J., Schulz, B. L., Cooper, C. A. et al., Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* 2004, *4*, 1650–1664.

[94] Ceroni, A., Maass, K., Geyer, H., Geyer, R. et al., GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.* 2008, *7*, 1650–1659.

[95] Gao, H. Y., Generation of asparagine-linked glycan structure databases and their use. *J. Am. Soc. Mass Spectrom.* 2009, *20*, 1739–1742.

[96] Ozohanics, O., Krenyacz, J., Ludanyi, K., Pollreisz, F. et al., GlycoMiner: a new software tool to elucidate glycopeptide

composition. *Rapid Commun. Mass Spectrom.* 2008, *22*, 3245–3254.

[97] Apte, A., Meitei, N. S., Bioinformatics in glycomics: glycan characterization with mass spectrometric data using SimGlycan. *Methods Mol. Biol.* 2010, *600*, 269–281.

[98] Gaucher, S. P., Morrow, J., Leary, J. A., STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal. Chem.* 2000, *72*, 2331–2336.

[99] Lapadula, A. J., Hatcher, P. J., Hanneman, A. J., Ashline, D. J. et al., Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. *Anal. Chem.* 2005, *77*, 6271–6279.

[100] Ethier, M., Saba, J. A., Spearman, M., Krokhin, O. 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*, 2713–2720.

[101] Shan, B., Ma, B., Zhang, K., Lajoie, G., Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *J. Bioinform. Comput. Biol.* 2008, *6*, 77–91.

[102] Tang, H., Mechref, Y., Novotny, M. V., Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* 2005, *21*(Suppl 1), i431–i439.

[103] Maass, K., Ranzinger, R., Geyer, H., von der Lieth, C. W. et al., "Glyco-peakfinder"–de novo composition analysis of glycoconjugates. *Proteomics* 2007, *7*, 4435–4444.

[104] Hahne, H., Kuster, B., A novel two-stage tandem mass spectrometry approach and scoring scheme for the identification of O-GlcNAc modified peptides. *J. Am. Soc. Mass Spectrom.* 2011, *22*, 931–942.

[105] Hayes, C. A., Karlsson, N. G., Struwe, W. B., Lisacek, F. et al., UniCarb-DB: a database resource for glycomic discovery. *Bioinformatics* 2011, *27*, 1343–1344.

[106] Marino, K., Bones, J., Kattla, J. J., Rudd, P. M., A systematic approach to protein glycosylation analysis: a path through the maze. *Nat. Chem. Biol.* 2010, *6*, 713–723.

[107] Herget, S., Ranzinger, R., Maass, K., Lieth, C. W., GlycoCT—a unifying sequence format for carbohydrates. *Carbohydr. Res.* 2008, *343*, 2162–2171.

[108] Ranzinger, R., Herget, S., Wetter, T., von der Lieth, C. W., GlycomeDB—integration of open-access carbohydrate structure databases. *BMC Bioinformatics* 2008, *9*, 384.

[109] Sahoo, S. S., Thomas, C., Sheth, A., Henson, C. et al., GLYDE—an expressive XML standard for the representation of glycan structure. *Carbohydr. Res.* 2005, *340*, 2802–2807.

[110] Campbell, M. P., Hayes, C. A., Struwe, W. B., Wilkins, M. R. et al., UniCarbKB: putting the pieces together for glycomics research. *Proteomics* 2011, *11*, 4117–4121.

[111] Hill, J. A., Smith, B. E., Papoulias, P. G., Andrews, P. C., ProteomeCommons.org collaborative annotation and project management resource integrated with the Tranche repository. *J. Proteome Res.* 2010, *9*, 2809–2811.

