A Review of Software Applications and Databases for the Interpretation of Glycopeptide Data

Matthew P. Campbell

Institute for Glycomics, Griffith University, Australia
FAX: +61–7–5552–9040, E-mail: m.campbell2@griffith.edu.au

(Received on April 25, 2017, accepted on May 31, 2017)

Key Words: glycoinformatics, glycoproteomics, database, data analysis, software

Abstract

Compared to proteomics, computational platforms for glycoproteomics is at an early stage and many researchers rely on the manual interpretation of large data sets to gain structural insights into the glycoproteome. Over the last few years there has been a steady increase in the availability of bioinformatics tools for processing and annotating glycoproteomics data sets. This mini-review describes advances in the development of algorithms and software applications and their applications. Furthermore, an update on structural and analytical databases is presented with a focus on those resources still actively maintained by the community, and how these resources are now being integrated into glycoproteomics pipelines to improve data interpretation.

A. Introduction

Protein glycosylation is by far the most important post-translation modification in terms of the number of proteins modified and the diversity generated. Nearly all membrane and secreted proteins as well as numerous intracellular proteins are modified by the covalent attachment of complex oligosaccharides (glycans) to protein backbones (1). Protein glycosylation is mediated by highly specific enzymes in the cellular machinery (2). This template-free biosynthetic apparatus generates glycan structural diversity comprising different monosaccharide compositions, anomericity/linkage types (α or β) and topologies (branched or linear) that reflect the physiology of the cell (3). In eukaryotes, glycosylation is usually divided into two main categories: N-glycosylation and O-glycosylation. N-glycosylation is largely restricted to the asparagine side chain in the consensus sequon Asn–Xxx–Ser/Thr (Xxx≠Pro) (4). In contrast, O-glycosylation shows more diverse forms in terms of the attachment site and linker monosaccharide. As described by Moremen et al., at least six distinct glycan types are known in animal systems (5). Most commonly monosaccharide residues N-acetylgalactosamine (GalNAc), N-acetylglucosamine (GlcNAc), xylose, mannose, fucose and glucose are attached to Ser and Thr. Alternatively, they can be attached through C–C linkages to Trp (C-mannosylation) or as a linker connecting glycosylphosphatidylinositol (GPI) anchors with protein backbones. Some specific types, such as the O-linked attachment of fucose and glucose are sequon dependent, whilst little is known on the complexities and biological roles of O-mannosylation (6–8). For a broader introduction to protein glycosylation, classes and types found in nature readers are referred to a number of comprehensive reviews (4, 5, 9–13).

By influencing the structure and function of proteins, glycosylation plays an essential role in a wide range of biological processes, such as intra- and intercellular signalling, developmental processes, tumour immunology, protein folding, and protein stabilization (12, 14–16). Aberrant protein glycosylation has been associated with many diseases such as pancreatic, breast, and gastric cancer (17), multiple sclerosis (18), inflammation (19–21) and Alzheimer’s disease (22). The comprehensive characterisation of protein glycosylation at high sensitivity is a major step for either the early detection of disease or the evaluation of therapeutic efficacy for treatment of disease.

To fully understand the functional roles of glycans and glycoproteins it is vital to gain an insight into the complete repertoire of oligosaccharides present. The accurate comparison of glycoforms and relative quantitation of oligosaccharides are necessary steps in this direction. Analytical techniques to identify glycan structures in a given sample and measure their abundance encompass several orthogonal methodologies including high-/ultra-performance liquid chromatography (H/UPLC), capillary electrophoresis (CE), nuclear magnetic resonance (NMR), mass spectrometry (MS) and tandem mass spectrometry (MS²) (23). Experimental factors governing quantification can be found in comprehensive reviews (24, 25) and recent work highlight the importance of using orthogonal techniques to fully determine and elucidate the abundance of glycan structures (26).

There are two main strategies for elucidating glycosylation information using MS techniques: (1) global characterization after the release of glycans from glycoproteins and (2) characterization of glycopeptides after proteolytic digestion. Glycoprofiling of released glycans is particularly useful when rapid analysis of glycan

© 2017 FCCA (Forum: Carbohydrates Coming of Age)
composition is desired, however, no information on protein attachment, i.e., site-specific glycosylation, can be obtained. To gain glycosylation site-specific information for individual glycoforms, the second method, glycopeptide analysis, which requires digestion of the protein using a protease such as trypsin is necessary. This method is generally advantageous because it provides information about both glycan composition and the site of the glycan’s attachment. However, the structural variation of glycosylation at any one site usually cannot be determined by high resolution MS data alone. It requires a combination of MS-based mapping and MS/MS sequencing of glycopeptides to generate sufficient information. This is less challenging when limited to the analysis of single or few glycoproteins of known identity, such that the m/z values of the proteolytically generated peptides can be predicted in silico for mass matching. As noted by Wu et al., there are instances when isobaric ambiguity, non-specific cleavages and additional peptide modifications occur, which prevent direct identification, therefore necessitating complete MS/MS sequencing (27). The technical challenges faced have been reviewed by others (28–30).

Compared with advances in instrumentation and experimental workflows, the development of software tools supporting (semi-)automated glycoproteomics analysis remains immature compared to computational proteomics (31). While the overall workflow for glycoproteomics bears similarity to those used in proteomics, major differences exist in the strategies for glycopeptide identification, validation of results and subsequent quantification of the identified species. As such there is an increasing demand for glycoinformatics solutions that can help address the various data processing and data interpretation challenges in the field. New bioinformatics tools and databases (both open access and commercial) are being developed at a rapid pace to analyze such datasets and to cope with new types of data. Unfortunately, most glycoproteomics software are designed to support the requirements and workflows of individual laboratories.

B. Glycomics and Glycoproteomics Databases

The purpose of this mini-review is to summarize databases and applications available to researchers using mass spectrometry workflows for analyzing glycopeptides, plus to provide an update on relevant glycomics (structural and experimental) databases. For a broader description of software applications used for the analysis of glycopeptide mass spectra, readers should refer to reviews (32–38). Readers interested in the use of peptomics for analysis of deglycosylated peptides are referred to the following review (39).

Table 1 lists relevant glycomics- and glycoproteomics-focused databases that are actively maintained and brings together descriptions summarized by other reviews (40–44). For the purposes of this mini-review we have included glycan structure databases, whose content is now increasingly integrated into glycoproteomic analysis tools (see Table 2).

The initial step toward site-specific characterization is to identify potential glycosylation sites. There has been a steady increase in the number of databases that provide knowledge of known or putative glycosylation sites, notably UniProtKB (45), UniCarbKB (46, 47), dbOGAP (48), dbPTM (49) and GlycoProtDB (50). Such databases can be divided into i) analytical databases that provide tools for querying and comparing data collections and ii) those tailored towards storing structure information and associated metadata, e.g. binding partners, biosynthesis pathways (glycosyltransferases and glycosidases), attached proteins and lipids, contextual disease information, and 3D modelling. In addition, established life science databases including UniProt, nextProt, and the Protein DataBank (PDB) provide glycan-related content.

UniProtKB and UniCarbKB both store information predominantly for N- and O-linked glycosylation sites; dbOGAP provides information on O-GlcNAc-modified proteins, and dbPTM (49) provides contextual content for C-glycosylated proteins (49). Another useful resource is GlycoProtDB (with almost 3000 entries) that provides data on the attachment sites of N-glycans obtained experimentally through the proteomic analysis of N-glycosylated glycoproteins in humans, mouse and nematode (50, 51).

UniCarbKB is a resource for mammalian glycoprotein and annotation data. The database provides information on the oligosaccharides characterized from a glycoprotein at either the global or site-specific level. This evidence is accumulated from a peer-reviewed and manually curated collection of information on oligosaccharides derived from membrane and secreted glycoproteins purified from biological fluids and/or tissues. In comparison, UniProtKB provides information on putative and experimentally verified glycosylation sites, but no implicit information is provided about glycan structures observed. Instead, these entries are linked to UniCarbKB and indicated using the terms ‘high mannose,’ ‘hybrid’ or ‘complex.’

For mammalian entries UniProtKB reports putative/predicted sites of N-glycosylation (containing the N-glycosylation consensus sequence) identified by the neural network predictor tool NetNGlyc (52). Although UniCarbKB stores information about fewer glycosylation sites it does describe glycoforms identified on these sites. For completeness, information associated with glycoenzymes is provided by the CAZy database (53).

C. Glycosylation Site Prediction and Modelling Tools

Mapping of a glycoproteome can be enhanced by understand-
Table 1. Databases relevant for interpreting glycomics and glycoproteomics data.

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Description and Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbohydrate Structure Database (CSDB) (89)</td>
<td><a href="http://csdb.glycoscience.ru">http://csdb.glycoscience.ru</a> Carbohydrate structure database merged from bacterial, archaeal, plant and fungal databases. Provides access to bibliographic, taxonomic, NMR spectroscopic information.</td>
</tr>
<tr>
<td>dbOGAP</td>
<td><a href="http://pir.georgetown.edu/staff/huz/hulab.html">http://pir.georgetown.edu/staff/huz/hulab.html</a> A literature based database of O-GlcNAcylated proteins and sites.</td>
</tr>
<tr>
<td>dbPTM</td>
<td><a href="http://dbptm.mbc.nctu.edu.tw">http://dbptm.mbc.nctu.edu.tw</a> dbPTM compiles information on protein post-translational modifications from 14 databases such as the catalytic sites, solvent accessibility of amino acid residues, protein secondary and tertiary structures, protein domains and protein variations.</td>
</tr>
<tr>
<td>ECODAB (Escherichia coli O-antigen database) (90)</td>
<td><a href="http://www.nevyn.organ.su.se/ECODAB">http://www.nevyn.organ.su.se/ECODAB</a> A database focused on O-polsaccharide structures of E. coli lipopolysaccharides, with supporting 1H and 13C NMR chemical shift data, flipase and polymerase sequences, and literature references.</td>
</tr>
<tr>
<td>EK3D (91)</td>
<td><a href="http://www.iith.ac.in/EK3D/">http://www.iith.ac.in/EK3D/</a> A curated database of K antigens from various E. coli serotypes plus 3D structures/model entries.</td>
</tr>
<tr>
<td>Glyco3D (92)</td>
<td><a href="http://glyco3d.cermav.cnrs.fr">http://glyco3d.cermav.cnrs.fr</a> A collection of databases that provide information on the three-dimensional features of monosaccharides, disaccharides, oligosaccharides, polysaccharides, glycosyltransferases, lectins, monoclonal antibodies against carbohydrates, and glycosaminoglycan-binding proteins.</td>
</tr>
<tr>
<td>Glycobase (Dublin) (93)</td>
<td><a href="http://www.nibrt.ie/glycobase/show_nibrt.action">http://www.nibrt.ie/glycobase/show_nibrt.action</a> Database focused on UPLC, HPLC, CE and RP-HPLC elution positions for over 600 glycan structures with links to taxonomy, glycoprotein and supporting literature.</td>
</tr>
<tr>
<td>Glycobase (Lille) (94)</td>
<td><a href="http://glycobase.univ-lille1.fr/base/">http://glycobase.univ-lille1.fr/base/</a> Structure database with supporting NMR data sets.</td>
</tr>
<tr>
<td>GlycoCD (95)</td>
<td><a href="http://glycosciences.de/glycocd/index.php">http://glycosciences.de/glycocd/index.php</a> Carbohydrate recognition and glycan information on carbohydrate related clusters of differentiation antigens (CD).</td>
</tr>
<tr>
<td>GlycoMob (96)</td>
<td><a href="http://www.glycomob.org">http://www.glycomob.org</a> Ion mobility collision cross section values for released glycans standards and their fragments.</td>
</tr>
<tr>
<td>GLYCOSCIENCE.DE (97)</td>
<td><a href="http://www.glycosciences.de">http://www.glycosciences.de</a> Provides access to a number of databases and bioinformatics tools that support structure, mass spectrometry, NMR and PDB querying, searching and analysis. Glycan and glycoprotein 3D structure generation.</td>
</tr>
<tr>
<td>GlyTouCan (98)</td>
<td><a href="http://www.glytoucan.org">http://www.glytoucan.org</a> International glycan sequence repository that assigns a unique accession number to any glycan structure, fragment, or composition, submitted by the community.</td>
</tr>
<tr>
<td>MonosaccharideDB (97)</td>
<td><a href="http://www.monosaccharidedb.org">http://www.monosaccharidedb.org</a> A comprehensive database of monosaccharides descriptions.</td>
</tr>
<tr>
<td>SugarBind (100)</td>
<td><a href="http://sugarbind.expasy.org">http://sugarbind.expasy.org</a> A curated database providing access to lectin adhesins of viral and bacterial pathogens and biotoxins, and known carbohydrate ligands.</td>
</tr>
<tr>
<td>The Consortium for Functional Glycomics (101)</td>
<td><a href="http://www.functionalglycomics.org">http://www.functionalglycomics.org</a> Provides access to MALDI-MS glycan spectra from mouse and human tissue, and an extensive library of glycan array and gene microarray data sets.</td>
</tr>
<tr>
<td>The Japan Consortium for Glycobiology and Glycotechnology Database (JCGGDB)</td>
<td><a href="http://www.jcggb.jp/index_en.html">http://www.jcggb.jp/index_en.html</a> JCCGB is a collection of specialized databases. The latest release provides access to: GGDB (GlycoGene Database), LiDB (Lectin Frontier Database), GlycoPOD (Glyco-Science Protocol Online Database), GlycoProtDB (GlycoProtein Database), GMDB (Glycan Mass Spectral Database), LipidBank, GlycoEpipte, GALAXY amongst others.</td>
</tr>
<tr>
<td>UniCarb KnowledgeBase (UniCarbKB) (47)</td>
<td><a href="http://www.unicarbkb.org">http://www.unicarbkb.org</a> A curated database of protein glycosylation (global and site-specific) reported in the scientific literature.</td>
</tr>
<tr>
<td>UniCarb-DB (102)</td>
<td><a href="http://www.unicarb-db.org">http://www.unicarb-db.org</a> An LC–MS/MS glycan fragmentation database.</td>
</tr>
<tr>
<td>Unipep (103)</td>
<td><a href="http://www.unipep.org">http://www.unipep.org</a> Database describing theoretical N-linked glycosylation sites derived from UniProt that have been mapped to LC-MS data.</td>
</tr>
</tbody>
</table>
ing important determinants of protein glycosylation at both the sequence and structure levels (e.g. local protein secondary structure, protein surface geometry and accessibility) (54, 55). In addition, to the databases listed are a growing number of applications that predict the likelihood of site-occupancy. The interplay between N-linked glycosylation sites and protein features was revealed by Petrescu and colleagues who analyzed the glycan-protein, and the peptide primary, secondary, and tertiary structures around N-glycosylation sites (54). With advances in data-mining and machine-learning algorithms several computational tools have been developed to improve the accuracy and prediction of N-/O- protein glycosylation. These tools include GlycoMine (56), NetNGlyc and NetOGlyc (57), EnsembleGly (58), and GPP (59), whose algorithms are based on sequence-derived features; for example, the physical/chemical properties of amino acids, position-specific scoring matrices, protein secondary structure, and functional annotations. Only a few tools, NGlycPred (60) and GlycoMine$^\text{struct}$ (61) systematically assess the importance of structural features (e.g. amino acid main and side chain accessibility, and side-chain depth index) in glycosylation site prediction. Indeed, GlycoMine$^\text{struct}$ is the only computational framework that assembles protein-sequence and protein-structural features for both N- and O-glycosylation-site prediction.

GlyProt (available on GLYCOSCIENCES.de) (62) and Glycoprotein Builder (part of the GLYCAM-Web resource (63)) allow researchers to model glycoproteins starting with a PDB file and attaching glycans to specific glycosylation sites (Table 2). By combining knowledge from the above databases, it is possible to model and probe the structural implications of site-specific glycosylation.

### D. N-glycopeptide Analysis

Over the last decade an increasing number of semi-automated tools have been developed to assist the interpretation of N-linked glycopeptide MS data, which provide compositions consistent with high resolution MS data. Such applications include GlycoMod (http://web.expasy.org/glycomod/) (64), GlycoX (65), GlycoPep DB (http://hexose.chem.ku.edu/sugar.php) and GlycoSpectrumScan (http://glycospectrumscan.org) (66).

GlycoMod is a widely-used tool that uses experimentally determined MS data to predict glycans released from a submitted peptide or protein sequence, the monosaccharide composition of free or derivatized N- and O-glycans, as well as the monosaccharide composition of glycans on glycopeptides from proteolytically digested proteins. The prediction is based on calculation of the combination of monosaccharide masses with some rules restricting unlikely combinations (43). GlycoSpectrumScan and GlycoPep DB address some of GlycoMod’s limitations including the handling of multiple charged precursors. GlycoSpectrumScan is designed to analyse LC-MS data of intact glycopeptides from proteolytic digests. The web application uses MS1 data to determine glycopeptide composition, along with relative distribution of glycoforms at each of the sites and requires users to input: (1) oligosaccharide compositions of the N- and/or O-linked glycans present in the sample and (2) in silico derived peptide masses of proteolytically digested proteins with a potential number of N- and/or O-glycosylation sites. A key feature of GlycoPep DB is a built-in database of glycoforms that have been previously identified in MS data, but this does restrict glycan matches to those matching database entries.

The availability of accurate glycan structure databases is important in delivering glycopeptide analysis tools. Most currently available glycan structure databases use their own proprietary structure representation schema and contain numerous annotation errors, and some contain chemically synthesised glycan structures not present in nature. These cause problems when such glycan

<table>
<thead>
<tr>
<th>Software Name</th>
<th>Description and Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly-Spec (104)</td>
<td><a href="http://www.glycam.org">http://www.glycam.org</a> A web-based tool for predicting glycan specificity by integrating glycan array screening data and 3D structure.</td>
</tr>
<tr>
<td>GLYCAM (63)</td>
<td><a href="http://www.glycam.org">http://www.glycam.org</a> A platform dedicated to simplifying the prediction of three-dimensional structures of carbohydrates and macromolecular structures involving carbohydrates.</td>
</tr>
<tr>
<td>GlycoPattern (105)</td>
<td><a href="http://glycopattern.emory.edu/">http://glycopattern.emory.edu/</a> A web-based tool to support the analysis of glycan array data for the Consortium for Functional Glycomics.</td>
</tr>
<tr>
<td>Privateer (106)</td>
<td>Software for the conformational validation of carbohydrate structures.</td>
</tr>
</tbody>
</table>
databases are used for the annotation or mining of data generated in the laboratory. To this end, a number of resources are being developed that provide access to curated collections of structural data (external literature, database references and user annotations) including Qrator (67), Carbohydrate Structure Database (CSDB), UniCarb and others listed in Table 1. GlyTouCan, a glycans structure repository, will facilitate software development, but entries are minimally internally checked for errors.

E. Glycopeptide MS/MS Analysis

Recognizing the need to automate the laborious process of interpreting glycopeptide MS and MS2 data and to facilitate site-specific glycosylation analysis, several (open access and commercial) applications have been developed including GlycoMiner (68), GlyPID (69), GlycoPep Grader (70), Peptoonist (71), GlyDB (72), GlycoPeptideSearch (73), GlycoPeptide Finder, SweetSEQer (74), SimGlycan (75), Medicel N-glycopeptide library (76), Pinnacle and Byonic.

Each of these tools differs by computational strategies and many have been designed to support specific workflows and/or limited datasets. As described by Liang et al., all share the same fundamentals: (i) accurate mass measurement in precursor MS spectra; (ii) deducing putative glycan composition and peptide backbone sequence from LC-MS/MS spectra; (iii) pattern matching the marker fragment ions for peptide backbone and glycan against theoretical spectra from protein and glycan database, or attempt de novo sequencing of the glycan moiety; and (iv) scoring (77). A brief description for a selection of tools listed in Table 3 follows.

Byonic™ is a PTM-centric search engine used in glycoproteomics workflows due to its ability to annotate and identify intact N- and O-linked glycopeptides from MS/MS spectra in addition to other peptide modifications. High resolution/mass accuracy LC-MS/MS data using one or multiple fragmentation modes usually serve as input. The application scores, ranks, and identifies glycopeptides by searching separate predefined or user-defined protein and glycan composition databases. For CID- and HCD-MS/MS, the assigned score takes into account the most common saccharide oxonium ions (i.e., for HexNAc m/z 138.05496 and 204.08665 and for NeuAc m/z 292.10269 and 274.09213) and the most likely glycopeptide fragments, such as the peptide plus core HexNAc (Y1 ion) and Fuc if core fucosylated. For a comprehensive description of glycopeptide fragmentation and oxonium ion decomposition readers are referred to the following (78–81).

GlycoMaster DB searches a protein sequence database and a glycan database to identify peptide-glycan pair(s) that match tandem mass spectra generated by HCD/ETD fragmentation. The application requires users to provide a list of protein sequences (glycosylated and unglycosylated) in the sample. GlycoMaster DB then filters MS/MS spectra based on the presence of oxonium ions and monosaccharide losses to remove non-glycosylated peptide tandem mass spectra. A built-in glycan database is then used to assign best fitting glycans to the MS/MS spectra. The glycan match works similarly to those used for peptide identification in which matches to theoretical product ion m/z values are assigned an award or penalty value. The product ions are scored for glycan sequence matches and peptide sequence matches to produce a raw score reflecting the sum of the individual peak scores.

GlycoPeptideSearch is an application for the determination of glycopeptide composition from collisional dissociation tandem mass spectra (82). Designed for analyzing purified glycoprotein samples GlycoPeptide Search uses GlycomeDB (83) and a user submitted peptide file to generate a list of glycopeptide matches based on known fragmentation rules. First, the algorithm searches for oxonium ions then checks for product ions that correspond to the peptide and N-glycan core-containing up to three monosaccharide residues. The mass of the glycan calculated from the MS/MS spectra is searched against GlycomeDB.

A recent tool is GlycoSeq that implements a heuristic iterated glycan sequencing algorithm coupled with prior knowledge for automated elucidation of glycan structures within a glycopeptide from a collision-induced dissociation tandem mass spectrum (84). GlycoSeq employs rules of glycosidic linkage as defined by glycan synthetic pathways to eliminate improbable glycan structures and build reasonable glycan trees. The tools employs a similar strategy to GlycoFragWork (85) but attempts to identify glycopeptides in isolated glycoproteins (or simple mixtures) from their CID spectra rather than complex samples acquired using different fragmentation methods e.g. HCD, CID and ETD.

GlyPID is designed to characterize N-linked glycopeptides through the combined use of MS1 and MS2 information extracted from LC-MS/MS data (69). One of the benefits to the method is that no prior knowledge of the potential glycosylation or identity of the glycopeptide is required. Instead, GlyPID clusters glycopeptides based on observed series of masses differing by monosaccharide units in the MS data. The algorithm assigns glycopeptide monoisotopic ions and charge states using a method to filter isotopic clusters for consistency. The algorithm scores each cluster of co-eluting glycopeptides using MS and tandem MS data. A limitation to the program is that when low resolution data is used, there is a significant increase in the number of false-positive glycopeptide identifications. The latest release of GlyPID uses an enhanced algorithm with a scoring function that works for high-energy induced collision dissociation (HCD) MS/MS data. In contrast SweetSEQer
Table 3. Software available for analyzing and annotating glycan and glycopeptide data.

<table>
<thead>
<tr>
<th>Software Name</th>
<th>Description and Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cartoonist (108)</td>
<td>Link not available Labels MS peaks with cartoons for N-linked oligosaccharides released from glycoproteins. Cartoonist uses isotope envelopes as a filter for mass matching.</td>
</tr>
<tr>
<td>GlycanAnalysis (109)</td>
<td><a href="http://www.shimadzu.co.jp/aboutus/ms_r/soft.html">http://www.shimadzu.co.jp/aboutus/ms_r/soft.html</a> Plugin for Mass + +. Can be used to annotate MS/MS spectra with structures registered in GlycomeDB and KEGG GLYCAN by matching a library of diagnostic ions.</td>
</tr>
<tr>
<td>GlycoFragWork</td>
<td><a href="http://darwin.informatics.indiana.edu/col/GlycoFragwork/">http://darwin.informatics.indiana.edu/col/GlycoFragwork/</a> Identifies intact glycopeptides from multiple pre-aligned LC-MS/MS datasets reporting mass, elution time and abundance information.</td>
</tr>
<tr>
<td>GlycoMaster DB (110)</td>
<td><a href="http://www-novo.cs.uwaterloo.ca:8080/GlycoMasterDB">http://www-novo.cs.uwaterloo.ca:8080/GlycoMasterDB</a> For N-linked glycopeptides. Searches a protein sequence database and a glycan structure database to find the best pair of peptide and glycan. The glycan component is only identified with HCD spectra.</td>
</tr>
<tr>
<td>GlycoMod (64)</td>
<td><a href="http://web.expasy.org/glycomod/">http://web.expasy.org/glycomod/</a> Given an experimentally determined mass (glycan and/or glycopeptide sequence) the tool determines all possible glycan or glycopeptide compositions.</td>
</tr>
<tr>
<td>GlycoPep DB (111)</td>
<td><a href="http://hexose.chem.ku.edu/sugar.php">http://hexose.chem.ku.edu/sugar.php</a> N-glycan compositional assignment of glycopeptides from MSn data. Uses a ‘Smart searching’ approach to assign glycan masses from in-house databases.</td>
</tr>
<tr>
<td>GlycoPep Detector (112)</td>
<td><a href="http://glycopro.chem.ku.edu/ZZKHome.php">http://glycopro.chem.ku.edu/ZZKHome.php</a> Determines N-linked glycopeptides composition from ETD data. Protein sequence and a set of theoretical glycans must be defined by the user.</td>
</tr>
<tr>
<td>GlycoPep Grader (70)</td>
<td><a href="http://glycopro.chem.ku.edu/GPGHome.php">http://glycopro.chem.ku.edu/GPGHome.php</a> Determines N-linked glycopeptide composition. Requires user to define protein sequence and a set of theoretical glycans. Decoy glycopeptide library used for generating FDRs.</td>
</tr>
<tr>
<td>GlycoPep ID (113)</td>
<td><a href="http://hexose.chem.ku.edu/predictiontable.php">http://hexose.chem.ku.edu/predictiontable.php</a> Identifies the peptide moiety of either sulfated, sialylated, or both sialylated and sulfated glycopeptides from CID spectra.</td>
</tr>
<tr>
<td>GlycoPeptide Finder (114)</td>
<td><a href="http://edwardslab.bmcb.georgetown.edu/trac/GlycoPeptideSearch/">http://edwardslab.bmcb.georgetown.edu/trac/GlycoPeptideSearch/</a> Lists potential amino acids and glycan compositions based on the mass of a putative glycopeptide, and whether it is N- or O-glycosylated. Uses deconvoluted and deisotoped CID/CAD/HCD spectra and filters for oxonium ions. Support for calculating FCR via a decoy strategy.</td>
</tr>
<tr>
<td>GlycoPeptideSearch (73)</td>
<td><a href="http://glycoinfo.eurocarbdb.tree/master/application/GlycoWorkbench">http://glycoinfo.eurocarbdb.tree/master/application/GlycoWorkbench</a> Can be used to identify N-glycan glycopeptide from CID spectra by matching peptide-glycan pairs from user defined peptide sequences and glycan structures stored in GlycomeDB.</td>
</tr>
<tr>
<td>GlyPepID (69)</td>
<td>Active link not available Performs glycan annotation and glycopeptide site identification from CID and HCD spectra.</td>
</tr>
<tr>
<td>GlycoProfileAssigner (115)</td>
<td><a href="https://bitbucket.org/fergaljd/glycoprofileassigner">https://bitbucket.org/fergaljd/glycoprofileassigner</a> Assigns structures to UPLC data by matching possible glycan structures to peaks by GU values. These assignments are refined by using information from sequential exoglycosidase digestions.</td>
</tr>
<tr>
<td>GlycoSeq (84)</td>
<td><a href="https://github.com/dbaileychess/CSMSL">https://github.com/dbaileychess/CSMSL</a> A heuristic iterated glycan sequencing algorithm coupled with prior knowledge for automated elucidation of glycan structures within a glycopeptide from collision-induced dissociation tandem mass spectrum.</td>
</tr>
<tr>
<td>GlycoWorkBench (116)</td>
<td><a href="https://github.com/glycoinfo/eurocarbdb/tree/master/application/GlycoWorkbench">https://github.com/glycoinfo/eurocarbdb/tree/master/application/GlycoWorkbench</a> Facilitates the annotation of mass spectrometry data by matching experimental data against theoretical MS/MS fragmentation.</td>
</tr>
<tr>
<td>GPQuest (118)</td>
<td>Spectral library matching algorithm for N-glycopeptides using HCD tandem MS based on presence of oxonium ions. Uses a library of deglycosylated peptides to classify and identify intact glycopeptide tandem mass spectra.</td>
</tr>
<tr>
<td>MAGIC (119)</td>
<td><a href="http://ms.iis.sinica.edu.tw/COMics/Software/MAGIC.html">http://ms.iis.sinica.edu.tw/COMics/Software/MAGIC.html</a> MAGIC can be used to identify intact glycopeptides from CID spectra. It extracts b/y ions to perform a database search, and uses mass shift look-up table (compiled from biosynthetic rules) to score glycan compositions.</td>
</tr>
<tr>
<td>MultiGlycan (120)</td>
<td><a href="http://darwin.informatics.indiana.edu/col/MultiGlycan/">http://darwin.informatics.indiana.edu/col/MultiGlycan/</a> Tool can be used to analysis MALDI and ESI spectra. It matches experimental isotopic envelopes with theoretical envelopes of N-glycans.</td>
</tr>
<tr>
<td>Peptoonist (121)</td>
<td>Link not available Similar to Cartoonist but optimized for glycopeptides.</td>
</tr>
</tbody>
</table>
uses a de novo type approach to evaluate glycopeptides. It automates the process of assigning product ions resulting from glycan dissociation, which permits rapid screening of mass spectra to identify glycopeptides. The tool uses a unique dynamic programming approach to link MS/MS peaks that are approximately separated by a predicted glycan mass-to-charge.

MAGIC (Mass spectrometry-based Automated Glycopeptide IdentifiCation platform) automates the analysis of glycopeptide MS/MS generated on QTOF instruments using collisional dissociation in an untargeted manner. A Y1-ion pattern matching method, which detects Y1- and Y0-ions requires neither a glycosylated protein sequence nor a glycan database, it generates in silico MS2 spectra that serve as input to a database search engine e.g. Mascot. The glycan composition is then assigned using the calculated glycan mass and a reference table of theoretical compositions. Recently, the developers of MAGIC launched three new modules packaged in MAGIC-web called MAGIC+, Reports Integrator and Glycan Search (86). MAGIC+ is designed to perform targeted glycopeptide analysis that allows users to upload their own protein sequence file to find glycopeptides in their data. Reports Integrator supports the integration of results from Mascot and MAGIC-web to generate a complete protein/peptide-glycan summary report. Glycan Search is an independent tool that allows users to find various glycans from the GlycomeDB database. 

Peptoonist, an extension of Cartoonist by Goldberg and colleagues, uses a proteomics database engine to assign unmodified peptides in a digest and then assigns glycopeptides via a heuristic scoring method that combines precursor and fragment ion information to annotate glycopeptides (71). Glycopeptide tandem mass spectra are scored against a glycopeptide search space constructed from the assumed peptide sequence and a set of biosynthetically possible N-glycans. Unfortunately, Peptoonist is not publicly available.

Sweet-Heart accepts low resolution, low mass accuracy ion-trap MS2/MS3 mass spectra data (27). The application implements a scoring scheme based on a supervised machine learning algorithm that identifies the glycosylation of intact N-glycopeptides, and then further deduces the mass values of respective peptide backbones for further sequencing by either targeted multi-stage mass spectrometry (MS³) or electron transfer dissociation experiments. Sweet-Heart requires no prior knowledge of glycan or peptide mass input and was shown to outperform currently available tools in its ability to process ion-trap–based CID data. This workflow utilizes the ability of the Thermo-Fisher Fusion Tribrid mass spectrometry to use HCD data to trigger CID in the ion trap and ETD.

Over the last few years a collection of tools has been developed by the Desaire group including GlycoPep Grader, GlycoPep Detector and GlycoPep Evaluator. GlycoPep Grader determines glycopeptide composition from tandem mass spectrometric data, specifically ion trap data and calculates theoretical glycopeptide compositions from a list of target glycoprotein sequences and a set of theoretical glycan compositions (70). It relies on the identification of dissociation patterns for high mannose, hybrid, and complex N-linked glycoprotein types, including patterns specific to fucose or sialic acid residues. The scoring algorithm scores potential candidate compositions of the same nominal mass against MS/MS data through evaluation of the Y(1) ion and other peptide-containing product ions, across multiple charge states. Candidate glycopeptide compositions are prepared using GlycoMod or GlycoPep DB that are scored against tandem MS data. The tool then generates an FDR value using a decoy database that consists of a set of glycopeptides with neutral mass values within 50 ppm of the measured accurate masses. In comparison GlycoPep Detector is designed for interpreting ETD tandem mass spectra. The algorithm calculates the theoretical m/z values for c-, z- and y-ions for

<table>
<thead>
<tr>
<th>Software Name</th>
<th>Description and Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pinnacle</td>
<td><a href="http://www.optystech.com">http://www.optystech.com</a> Commercial software. Performs qualitative scoring on MS and combined MS/MS data while quantifying all precursor data to build comprehensive glycopeptide lists, and scores these using a multivariate scoring system. Software supports multiple fragmentation methods.</td>
</tr>
<tr>
<td>SimGlycan (75)</td>
<td><a href="http://www.premierbiosoft.com/glycan/">http://www.premierbiosoft.com/glycan/</a> Commercial software. Capability to process and interpret MSn data. Annotates experimental MS data by matching against an internal database and generates a scored list of candidate structures.</td>
</tr>
<tr>
<td>Sweet-Heart (27)</td>
<td><a href="http://sweet-heart.glycoproteomics.proteome.bc.sinica.edu.tw/">http://sweet-heart.glycoproteomics.proteome.bc.sinica.edu.tw/</a> Supervised machine learning to identify N-glycopeptides from ion trap based LC–MS/MS.</td>
</tr>
<tr>
<td>SweetSEQer</td>
<td><a href="http://software.steenlab.org">http://software.steenlab.org</a> Open source tool for identifying potential glycopeptide MS/MS spectra.</td>
</tr>
</tbody>
</table>

Table 3. Continued.
each candidate glycopeptide, and searches against the ETD tandem mass spectra and assigns a final score. Finally, GlycoPep Evaluator was developed to overcome problems using target-decoy libraries. Instead, the program calculates 20 mock glycopeptide compositions with masses isobaric to the true glycopeptide and uses these to estimate the false discovery rate for ETD tandem mass spectra.

F. Discussion

Site-specific characterization of protein glycosylation is essential for understanding the functional role of glycosylation in health and disease. The field of glycoproteomics is developing rapidly and improvements in analytical workflows and instrumentation are now generating large datasets.

In the last few years a number of biocuration activities have extended the coverage of protein glycosylation knowledge in databases, notably UniCarbKB, UniProt and nextProt. Cross-referencing between these resources has improved data connectivity and sharing that is assisting glycobiology and glycoproteomics analysis, but there are discrepancies in data content between these platforms. For example, as noted by Chalkey and Baker, UniProtKB reports five N-glycosylation sites for glucosylceramidase (two based on experimental data; Asn165 and Asn289) and three based on the presence of the consensus sequence motif, but UniCarbKB only stores information for Asn165 (the only site with determined glycans). This highlights the information gap problem and deficiencies in data presentation in these platforms, however, initiatives to improve database connectivity via Semantic Web technologies (e.g. GlycoRDF) aim to improve this situation.

In addition, Chalkey and Baker identified that neither of these databases provide information on O-glycosylation of Prolow-density lipoprotein receptor-related protein 1, Nucleobindin 1, and Fibronectin (87), which demonstrates the importance and impact of glycoscience biocuration and data sharing to increase protein glycosylation coverage in public databases.

A major limiting factor is data handling and analysis, which is a time- and labor-intensive process, requiring a high level of expertise due to the structural complexity of N-glycopeptides. Consequently, the design and development of bioinformatics and computational tools to assist and (semi-)automate glycopeptide identification is necessary. In this review, we have summarized the availability of current software applications and relevant databases used for site-specific glycosylation analysis of glycoproteins and glycoprotein mixtures. As discussed by Dallas and colleagues a major problem impacting the development of glycomics and glycoproteomics software is the lack of open source programs, which limits activities to extend functionality of existing tools and databases (34). Software quality assurance along with compliance with and promotion of good development practice are powerful mechanisms for stimulating software sustainability. However, systematic verification and validation of bioinformatics software is difficult. As highlighted by Hu et al. (88), it would be beneficial to have integrated and automated frameworks for glycoproteomics that provide common modules/libraries for file conversion, spectra preprocessing, sequence identification, quantification, and database connectivity, etc. ultimately leading to robust and scalable software solutions reducing the current bottlenecks of glycoproteomics strengthening its integration into glycoscience biocuration and data sharing to increase protein glycosylation coverage in public databases.

Acknowledgments

The author acknowledges funding support from the Institute for Glycomics and Griffith University.

References


Assignment of Saccharide Identities through Analysis of Oxonium Ion Fragmentation Profiles in LC MS/MS of Glycopeptides. *J. Proteome Res.* **13**, 6024–6032.


**Information of the Authors**

Dr. Campbell has had a leading role in the development of pioneering frameworks and informatics solutions, in both the academic and industrial sector, in the glycoinformatics area that have included EUROCarbDB, GlycoBase and the UniCarb initiative. His research career commenced at the Oxford Glycobiology institute under the supervision of Prof. Pauline Rudd before relocating to the National Institute for Bioprocessing Research and Training (NIBRT, Dublin, Ireland). In 2010 he joined Prof. Nicolle Packer’s group (Macquarie University, Sydney, Australia) to establish the international UniCarb program supported by the Australian National eResearch Collaboration Tools and Resources project (NeCTAR) and European funding agencies. In January 2017 Dr Campbell relocated to the Institute for Glycomics, a flagship biomedical research institute at Griffith University’s Gold Coast campus to start up a glycoinformatics research program. The Institute for Glycomics represents a unique national resource with a multi-disciplinary and translational approach with a particular focus in the area of glycoscience and the application of “omics” approaches to the study of carbohydrates in biological systems. It is one of a few such institutes in the world and the largest in the southern hemisphere and is now home to the newly-established Australian Centre for Cancer Glycomics.