Chapter 18

Databases and Associated Tools for Glycomics and Glycoproteomics

Frederique Lisacek, Julien Mariethoz, Davide Alocci, Pauline M. Rudd, Jodie L. Abrahams, Matthew P. Campbell, Nicolle H. Packer, Jonas Ståhle, Göran Widmalm, Elaine Mullen, Barbara Adamczyk, Miguel A. Rojas-Macias, Chunsheng Jin, and Niclas G. Karlsson

Abstract

The access to biodatabases for glycomics and glycoproteomics has proven to be essential for current glycobiological research. This chapter presents available databases that are devoted to different aspects of glycobioinformatics. This includes oligosaccharide sequence databases, experimental databases, 3D structure databases (of both glycans and glycorelated proteins) and association of glycans with tissue, disease, and proteins. Specific search protocols are also provided using tools associated with experimental databases for converting primary glycoanalytical data to glycan structural information. In particular, researchers using glycoanalysis methods by U/HPLC (GlycoBase), MS (GlycoWorkbench, UniCarb-DB, GlycoDigest), and NMR (CASPER) will benefit from this chapter. In addition we also include information on how to utilize glycan structural information to query databases that associate glycans with proteins (UniCarbKB) and with interactions with pathogens (SugarBind)

Key words Glycobioinformatics, Oligosaccharide, Omics, HPLC, NMR, Mass spectrometry, Databases

1 Introduction

The success of bioinformatics in providing breakthroughs in life science research has been gradual, and it is hard to say when it all began. Publically accessible life science databases, with their symbiotic connection to the open Internet, turned the concept into reality at the end of the last century. At the same time, the need to query databases with intelligent software increased. Genomic databases are currently the dominating resource, and the population of these databases with genome sequences has been boosted by

Electronic supplementary material: The online version of this chapter (doi:10.1007/978-1-4939-6493-2_18) contains supplementary material, which is available to authorized users.

Gordan Lauc and Manfred Wuhrer (eds.), *High-Throughput Glycomics and Glycoproteomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1503, DOI 10.1007/978-1-4939-6493-2_18, © Springer Science+Business Media New York 2017

analytical developments starting from Sanger sequencing to what is now known as NextGen sequencing that allows sequencing of the genome (DNA) and transcriptome (mRNA). These nucleotide databases (GenBank, European Nucleotide Archive, and DNA Data Bank of Japan) are now invaluable resources for the life sciences. Subsequently, with the improvement in biological mass spectrometric ionization techniques such as ESI and MALDI, large-scale protein identification (proteomics), utilizing nucleotide databases, became possible. Other -omics such as metabolomics and, as is the case for this chapter, glycomics/glycoproteomics have been more challenging because the molecules are secondary gene products that cannot be predicted from the genomic databases. Instead, the generation of glycan structural databases relies on data from the glycoanalytical field, and on the development of glycobioinformatics standards and symbolic nomenclature to allow the transition from the glycocommunity to the wider life science arena. Early attempts to provide glycostructural databases include CarbBank [1], EUROCarbDB [2], and GlycoSuiteDB [3], and their legacy whereby they are now integrated into the UniCarb family [4, 5], GlycomeDB [6], GlycoBase [7], and GRITS (www.grits-toolbox.org). With the realization that understanding post-translational modifications is key to describing the functional molecules that operate in every cell in an organism, glycodatabases and associated tools have become a prerequisite for life science research, not only for addressing glycobiology specifically, but for overall understanding of cell systems biology. The integration of various -omics data will be required for modulating an in silico cell (Fig. 1).

Glycomics databases that are available today can roughly be divided into analytical databases that can be queried and compared with data to predict structures, and databases that can be queried by the resultant glycan structural information to obtain associated metadata (including binding partners, associated glycosyltransferases and glycosidases, attached proteins, lipids, related organisms and diseases, and prediction of 3D structures). Glycosylation information can also be found in more generic life science databases such as UniProt, Protein DataBase, and Protein Atlas. Glycomics has been provided with a separate tab in ExPASy (www.expasy.org/glycomics) that includes both internal and external links to databases and tools. Without claiming to describe every resource that is available we provide a table (Table 1) that contains many glycodatabases. In addition we also describe in detail a selection of tools that can be used to interpret glycomic data generated by well-established analytical methods, including MS, NMR, and HPLC. The focus of this chapter is on the application of bioinformatics to the *N*-linked and *O*-linked glycans on mammalian glycoproteins. This chapter also includes examples of how to indicate the way glycan structures interact with pathogens, and how to find proteins on which a particular glycan structure(s) is expressed. The methods described here are devoted to interpretation of data generated

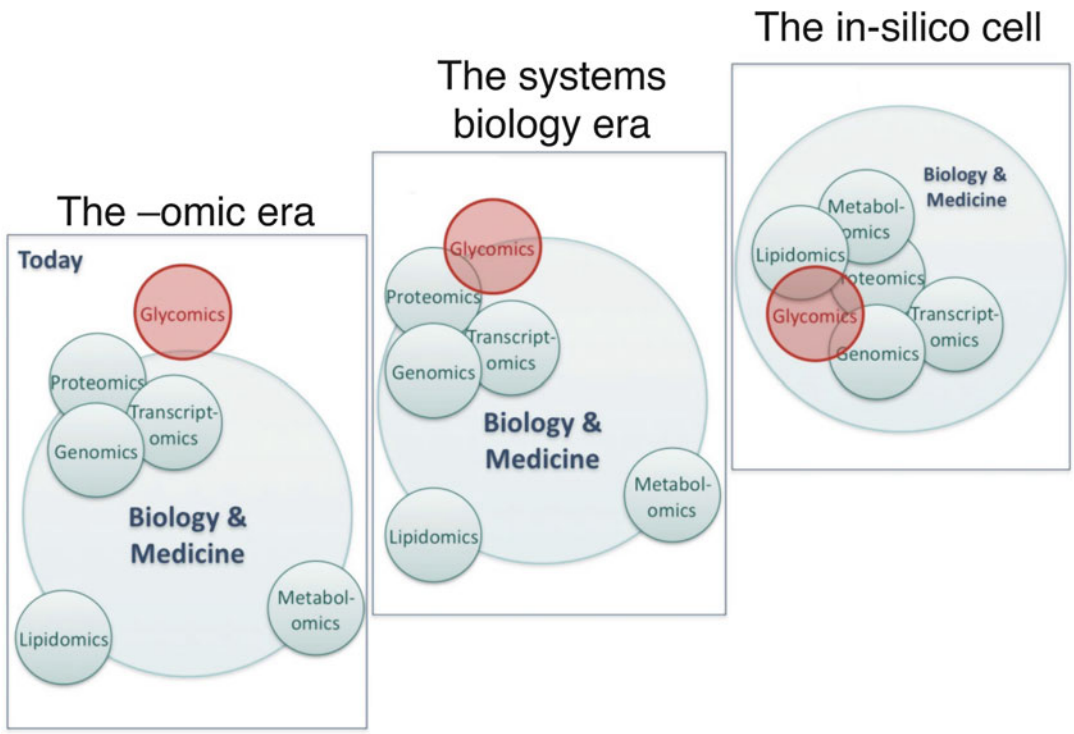


Fig. 1 The future of integration of -omics research in biology and medicine in order to understand the interplay between molecules of life

by glycomic/glycoproteomic analyses. There are other chapters and literature (e.g., Hayes et al. [8] and Campbell et al. [9]) that specifically address the workflows for generation of data that can be used for the tools presented in this chapter. We provide datasets as supplementary material to this chapter for the reader to test the different query functions of databases and tools described herein.

2 Materials

2.1 From Analytical MS Data to Monosaccharide Composition: GlycoMod

GlycoMod uses experimentally determined MS data to predict, for glycans released from a submitted peptide or protein sequence, the monosaccharide composition of free or derivatized *N*- and *O*-glycans (*see* **Note 1**), 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 [10].

Online website: web.expasy.org/glycomod/
Type of data: MS peak list can be submitted from both negative and positive ion mode mass spectrometry.

Table 1
List of databases relevant for glycomics and glycoproteomics

Science databases that inclusively contain glycoinformation
The Universal Protein Resource (*UniProt*) is a resource for protein sequence with annotations, www.uniprot.org. *Glycodatabase related*: Predicted and confirmed sites of glycosylation within protein amino acid sequences. *References*: Farriol-Mathis et al. [40]
The Protein Data Bank (*PDB*) is a database containing three-dimensional data from large biological molecules, such as proteins and nucleic acids, www.rcsb.org. *Glycodatabase related*: Information about localization of *N*-and *O*-linked glycosylation sites as well as ligand localization of carbohydrate-binding proteins. *References*: Berman et al. [41]
The Gene Expression Omnibus (*GEO*) gene expression data repository, www.ncbi.nlm.nih.gov/geo. *Glycodatabase related*: Expression profiles of glycorelated genes in cell models, animal models, and clinical samples. *Reference*: Barett et al. [42]
The Human Proteome Atlas is a tissue, cell line, and cancer tissue map of the proteome, www.proteinatlas.org. *Glycodatabase related*: Glycoprotein and glycogene localization in tissues, as well as subcellular localization. *References*: Uhlén et al. [43]
LipidBank is a database of natural lipids, www.lipidbank.jp. *Glycodatabase related*: Literature data and experimental evidence (e.g., MS, UV, IR, NMR) of glycolipids. *References*: Watanabe et al. [44]
LIPID MAPS contains chemical representation and annotations of biologically relevant lipids, www.lipidmaps.org. *Glycodatabase related*: Chemical representations of glycoplipids from consortium and literature data. *Reference*: Sud et al. [45]
Exclusive glycodatabases
The Consortium for Functional Glycomics (*CFG*) is an international initiative providing glycomic services and databases, www.functionalglycomics.org. *Glycodatabase related*: Data from the screening of the consortium's glycan array platform, glycogene microarray of tissues and cells, MALDI-MS screening of glycans from mouse and human tissue. It also contains information about histology, immunology, hematology, and metabolism/behavior of glyco-related gene mouse knockouts. The "CFG Molecule Pages" integrate data generated from the CFG screening with public data about glycosyltransferases, glycan structures, and glycan-binding proteins. Glycan array data from the consortium can be mined by GlycoPattern (glycopattern.emory.edu) [46].
The UniCarb KnowledgeBase (*UniCarbKB*) is a curated database of information on the glycan structures of glycoproteins reported in scientific literature, www.unicarbkb.org. *Glycodatabase related*: Current version contains 3238 structures from the literature, including information about taxonomy, tissue, and associated protein. *Reference*: Campbell et al. [29]

(continued)

Table 1
(continued)

The UniCarb-DB database is an assembly of glycosequences with associated MS fragmentation data, www.unicarb-db.org. *Glycodatabase related*: Current version contains more than 600 spectra from published work, including information about structure, taxonomy, tissue, and associated protein. *Reference*: Hayes et al. [4]
GlycomeDB is a metadatabase of carbohydrate sequences, www.glycome-db.org. *Glycodatabase related*: Database includes structures from CFG, KEGG, GLYCOSCIENCES.de, BCSDB, and CarbBank [1], as well as information about species and associated aglycon. *References*: Ranzinger et al. [6]
GlyTouCan, International glycan sequence repository, www.glytoucan.org. *Glycodatabase related*: Allows registration of a glycan structure with minimal metadata to obtain a unique accession number. *References*: Aoki-Kinoshita et al. [47]
Kyoto Encyclopedia of Genes and Genomes (*KEGG*) is a database of genome sequences and other high-throughput data, www.genome.jp/kegg. *Glycodatabase related*: www.genome.jp/kegg/glycan. Pathways for glycan biosynthesis and metabolism, roles of glycans in cellular processes, glycans in cancer, glycosyltransferase reactions and relation to diseases, carbohydrate binding proteins, The KEGG GLYCAN structure database contains structures from CarbBank [1], literature, and from structures present in KEGG pathways. *Reference*: Kanehisa et al. [48], Hashimoto et al. [49]
The Japan Consortium for Glycobiology and Glycotechnology DataBase (*JCCGB*) is a metadatabase that allows one to simultaneously search in glyco related databases, www.jcggdb.jp. *Glycodatabase related*: The current version includes GGDB (GlycoGene Database, www.jcggdb.jp/rcmg/ggdb), LfDB (Lectin Frontier Database, www.jcggdb.jp/rcmg/glycodb/LectinSearch), GlycoPOD (Glyco-Science Protocol Online Database, www.jcggdb.jp/GlycoPOD), GlycoProtDB (GlycoProtein Database, www.jcggdb.jp/rcmg/gpdb), GMDB (Glycan Mass Spectral Database, www.jcggdb.jp/rcmg/glycodb/Ms_ResultSearch), LipidBank, GlycoEpitope, GALAXY (Glycoanalysis by the three axes of MS and chromatography), KO mice DB (Functional Glycomics with KO Mice Database, www.metadb.riken.jp/metadb/db/Glycomics_mouse), Glycoforum, PACDB (Pathogen Adherence to Carbohydrate Database, www.jcggdb.jp/search/PACDB.cgi), GDGDB (Glyco-Disease Genes Database), TuMaRdb (Tumor Markers Reference Database), JCGGDB Reports, GlycosidaseDB, and JIGSDB (JCGG Integrated Glycan Structures Database). *Reference*: Maeda et al. [50]
The GlycoEpitope is a database of 173 common names of glycobiological epitopes associated with known oligosaccharide sequence, www.glycoepitope.jp.
GALAXY is a database of glycan sequences. Glycoanalysis using MS and chromatography, www.glycoanalysis.info. *Glycodatabase related*: Elution profiles of PA-labeled *N*-linked oligosaccharides analyzed by different LC-methods. *Reference*: Tomiya et al. [51]
GlycoDomainViewer is a graphic interface that displays a database of *O*-linked and *N*-linked glycosylation sites on proteins which are based on UniProt annotation and in-house discovery, glycodomain.glycomics.ku.dk. *Glycodatabase related*: Database of *O*-linked sites has been used in NetOGlyc neural network predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins (www.cbs.dtu.dk/services/NetOGlyc). *Reference*: Steentoft et al. [52]

(continued)

Table 1
(continued)

GLYCOSCIENCES is a database of carbohydrate sequences from literature that provides taxonomy, linking to PDB, NMR and 3D modules, www.glycosciences.de. *Glycodatabase related*: Database built on CarbBank [1] which is updated by including glycans from PDB. It also includes the MonosaccharideDB, a database of monosaccharides (www.monosaccharidedb.org). *Reference*: Lütteke et al. [53]
Glycobase (*Lille*) is a glycan sequence database which includes taxonomy and NMR data, glycobase, www.glycobase.univ-lille1.fr/base.
Glycobase (*Dublin*) is a database of glycan sequences that includes experimental HPLC data, taxonomy, MS and literature references. glycobase, www.nibrt.ie/glycobase/show_nibrt.action. *Glycodatabase related*: Current version v3.2 include 655 sequences. *Reference*: Campbell et al. [7]
Carbohydrate Structure Database (*CSDB*) is a merged database of Bacterial (BCSDB), and Plant and Fungal (PFCSDB) databases, www.csdb.glycoscience.ru. *Glycodatabase related*: Structures of glycans and NMR data from literature.
Glyco3D is a portal for databases covering three-dimensional features of monosaccharides, disaccharides, oligosaccharides, polysaccharides, glycosyltransferases, lectins, monoclonal antibodies against carbohydrates, and glycosaminoglycan-binding proteins, www.glyco3d.cermav.cnrs.fr/home.php. *Glycodatabase related*: The database contains three-dimensional data on oligosaccharides (BiOligo), lectins (Lectin3D), glucosaminoglycan-binding proteins with ligands (GAG), monoclonal antibodies with carbohydrate ligands (mAbs), glycosyltransferases (GT), and polysaccharides (Polysac). *Reference*: Pérez et al. [54]
The Carbohydrate-Active enZYmes Database (*CAZy*) contains genomic, structural and biochemical information on glyco-enzymes, www.cazy.org. *Reference*: Lombard et al. [55]
The Unipep database contains theoretical N-linked glycosylation sites in UniProt that have been mapped to LC-MS data, www.unipep.org. *Reference*: Zhang et al. [56]
GLYCAM libraries contain predicted 3D structures of oligosaccharides. Structures are available for *N*-linked, *N*-acetyllactosamine building blocks, and from the CFG-array, glycam.org/PreBuiltLibraries.jsp. *Reference*: Kirschner et al. [57]
SugarBind Database contains literature references on lectin adhesins of viral and bacterial pathogens and biotoxins, and their known carbohydrate ligands, www.sugarbind.expasy.org. *Reference*: Mariethoz et al. [32]
ECODAB (*Escherichia coli* O-antigen database) contains O-polysaccharide structures of *E. coli* lipopolysaccharides, putative, predicted, or confirmed glycosyltransferase functions, ^{1}H and ^{13}C NMR chemical shift data, flippase and polymerase sequences, and literature references, www.nevyn.organ.su.se/ECODAB. *Reference*: Rojas-Macias et al. [22]
EK3D is a manually curated database of K antigens corresponding to various *E. coli* serotypes, and their modeled 3D structures, www.iith.ac.in/EK3D. *Reference*: Kunduru et al. [58]

Test data for:

N-linked glycopeptides: *N*-linked glycopeptide of human alpha-HS-glycoprotein (P02765), reduced and alkylated with iodoacetamide. $[M+H]^+$ ion of m/z 3394.4495 (±0.05 Da)

Reduced, released *O*-linked oligosaccharides: $[M-H]^-$ ions of m/z 384.1, 675.1, 966.2, and 1331.2 (±0.3 Da)

Released AB-labeled *N*-linked glycans: $[M-H]^-$ ions of m/z 1581.92, 1744.00, and 1906.01 (±0.5), mass of 2-aminobenzamide 136.06 Da

2.2 From Structure to Predicted MS/MS Data: GlycoWorkbench

GlycoWorkbench is designed to facilitate the annotation of mass spectrometry data while presenting corresponding glycan structures [11, 12]. A particular feature of this software is that it provides a friendly interface for drawing various glycans of different types, labeling, and charge states. The theoretical MS/MS fragment of a selected structure can be computed.

Website for downloading tool: code.google.com/p/glycoworkbench/ (static); github.com/alternativeTime/glycoworkbench (current)

Manual: http://bit.ly/1Q3pAiI

Type of data: MS/MS peak lists of glycans

Test MS/MS data: A peak list of a PA-labeled *N*-linked glycan with $[M-H]^-$ ion of m/z 1311.08 and a composition of five hexoses and two *N*-acetylhexosamines (Supplementary Table 1).

2.3 From Analytical MS and MS/MS Data to Structure: UniCarb-DB

UniCarb-DB [4] provides access to a collection of LC-MS/MS *N*- and *O*-linked glycan fragments released from glycoproteins and complemented with experimental metadata. UniCarb-DB supports the comparison of users' experimental spectra against the experimental spectra in the database. Therefore, UniCarb-DB can be used as reference to aid manual annotation and development of more precise software tools to analyze and interpret spectrometric data.

Online website: www.unicarb-db.org

Type of data: MS/MS fragment peak list from bar spectra together with pseudomolecular ion mass (*see* **Note 2**)

Test MS data for MS/MS search: MS/MS peak lists of a reduced *N*-linked oligosaccharide with [M−H] ion of m/z 1056.48 (Supplementary Table 2).

2.4 From NMR Data to Structure: CASPER

The computer program CASPER (Computer Assisted SPectrum Evaluation of Regular polysaccharides) was originally developed to facilitate structural elucidation of polysaccharides using input sugar

and methylation analyses together with *unassigned* ^{13}C NMR data [13]. Subsequently, ^{1}H and 2D NMR data were made possible as input to the program [14–16] (which was also enhanced to include oligosaccharides [17]), in the analysis of unknown structures. CASPER is freely available online at the website www.casper.organ.su.se/casper/ [18].

CASPER is used in two modes: (1) prediction of ^{1}H and ^{13}C NMR chemical shifts using a given structure as input or (2) structural determination of a glycan using unassigned ^{1}H and/or ^{13}C NMR data as well as components [19, 20] and linkage positions when known, e.g., as suggested by bioinformatics information in the database ECODAB [21, 22]. Besides NMR chemical shift predictions and structural elucidation based on the constituent sugar residues, substituents such as *O*-acetyl, phosphoryl, sulfuryl, and pyruvyl groups can be added. Glycosides functionalized at the reducing end by amino acids, methyl, propyl, 2-azidoethyl, and trifluoroacetyl-*para*-aminophenylethyl groups are also available as well as sugar residues reduced to their corresponding alditols [23, 24].

2.5 From HPLC/UHPLC/CE Data to N-linked Structure: GlycoBase

The National Institute of Bioprocessing Research and Training (NIBRT) experimental database, GlycoBase is an experimental resource that allows users to conduct structural assignments based on standardized liquid chromatography retention times, expressed as glucose unit (GU) values. Further confirmation of structures can be achieved using exoglycosidase digestions followed by further chromatographic analysis to determine monosaccharide sequence and linkage.

On-line open access website: glycobase.nibrt.ie

Type of data: Chromatographic data: Released glycans labeled with 2-aminobenzamide with GU values and relative percent areas. A description of how data can be generated is found in Chapters 4 and 10 and in references [8, 25–27]. Test data can be found in the table in Fig. 7. 8-aminopyrene-1,3,6-trisulfonic acid-labeled glycans can be analyzed on CE using the same rationale.

2.6 Validation of Monosaccharide Sequence and Linkages with the Aid of Exoglycosidase Digestion: GlycoDigest

One of the most commonly used methods for determining the sequence and structure of oligosaccharides is the enzymatic analysis of oligosaccharides using highly specific exoglycosidases, either sequentially or in array form. GlycoDigest is a web-application and standalone tool that simulates the behavior of these exoglycosidase digestions on released *N*- and *O*-linked oligosaccharides. The tool is based on a set of rules acquired from expert knowledge and supporting experimental evidence provided by NIBRT's GlycoBase, and can be used to tailor a panel of exoglycosidases to unambiguously confirm oligosaccharide structure.

Online website: www.glycodigest.org

Type of data: A description of the enzymes supported and a user guide can be accessed at www.glycodigest.org and in reference [28].

2.7 Literature Report on Structure and Association with Protein: UniCarbKB

UniCarbKB is an international collaboration that develops open source glycobioinformatics databases to support glycomics and the emerging technology of glycoproteomics [29]. It is a curated collection of information on well-characterized glycan structures derived from glycoproteins, biological fluids, and tissues. Although UniCarbKB provides annotated entries for many species, its primary focus is on proteins from mammalian systems of distinct taxonomic groups to ensure the presence of high quality annotations for representative members of all protein families. The database provides contextual information for *N*- and *O*-glycan structures attached to proteins and where known, maintains the commonly lost connection between a glycan structure and its attached protein(s), as annotated in UniProtKB [30] and neXtProt [31]. In addition, an entry is supplemented with descriptions of its biological source, supporting reference and experimental methods employed to determine the glycan structure.

Online website: www.unicarbkb.org.

2.8 Association of Structure with Microbial Interactions: SugarBindDB

SugarBindDB is a curated database that covers knowledge of glycan binding of human pathogen lectin adhesins. A set of three inseparable components constitutes the core information that is stored, i.e., the pathogenic agent, a lectin adhesin, and a glycan ligand. Each of these entities is named with as much precision as possible: taxonomic designation for pathogen agent, protein name for lectin, and epitope name for ligand. Furthermore, each glycan-binding event is backed with at least one published reference. SugarBindDB is cross-linked to protein and glycan-related resources such as UniProtKB and UniCarbKB. It is tightly bound to the latter via a substructure search tool that maps each ligand to full structures when it is available. Other search, navigation, and visualization tools are implemented to consult, compare, and bring out information on glycan binding. For instance, SugarBindDB can be used to identify protein partners interacting via glycans as shown in Subheading 3.8.

Online website: sugarbind.expasy.org [32]

3 Methods

3.1 From Analytical MS Data to Monosaccharide Composition: GlycoMod

The most straightforward use of MS analysis for glycoconjugates is to determine monosaccharide compositions based on experimentally determined masses; and GlycoMod can be used for this purpose on both released glycans and glycopeptide MS data. GlycoMod is also cross-linked to UniCarbKB where predicted compositions

are linked to possible glycan sequences reported in the literature and found in the database. This allows the combinatorial possibilities of monosaccharides to be narrowed down to previously seen and reported structures in biological systems.

3.1.1 Searching GlycoMod Using Oligosaccharides and Glycopeptides

1. Launch the GlycoMod website (Fig. 2, numbers below refer to this figure)
2. Export peak lists of singly charged pseudomolecular ions from mass spectrometric software (*see* **Note 3**).
3. Paste the peak list of pseudomolecular ions (1) together with information on type of adducts and ion mode (2) and MS accuracy (3) into the website (*see* test data in Subheading 2.1).
4. Enter information into the website about the type of glycoconjugate you are analyzing:
   (a) *N*- or *O*-linked
   (b) Glycopeptides, free oligosaccharides, reduced oligosaccharides (4) (for the *O*-linked test data) and derivatized oligosaccharides (for the *N*-linked test data)
   (c) Glycoprotein sequence or UniProt ID, type of protease, and cysteine modification (5) (for glycopeptide test data insert P02765, trypsin and iodoacetamide, respectively)
   (d) Reducing end derivatization mass (6) (for the *N*-linked test data 136.06)
   (e) Type of monosaccharides (7) (*see* **Note 4**)
5. Submit the search
6. For each match, compositions are displayed that match submitted data. UniCarbKB database can be used initially to narrow down the possibilities. The data will be provided as compositions along with MS error and masses of displayed glycoforms (*see* **Note 5**).
7. For glycopeptide data, the result should be: 1736.65 (glycoform mass), −0.003 (Δmass (Dalton)), $(HexNAc)_2$ (Deoxyhexose)$_3$ + $(Man)_3(GlcNAc)_2$ (structure), 1656.796 (peptide mass [M]), and VCQDCPLLAPLNDTR (peptide sequence)
8. For the AB-labeled *N*-linked test data the compositions with UniCarbKB entries should be according to Table 2.
9. For the reduced *O*-linked test data, the compositions with UniCarbKB entries (in addition to compositions that are less likely to occur in nature) should be according to Table 3.

3.2 From Structure to Predicted MS/MS Data: GlycoWorkbench

GlycoWorkbench is designed for rapid structure drawing, theoretical fragment computation, and association of theoretical fragments with experimental MS/MS peak lists, as well as other many useful functions.

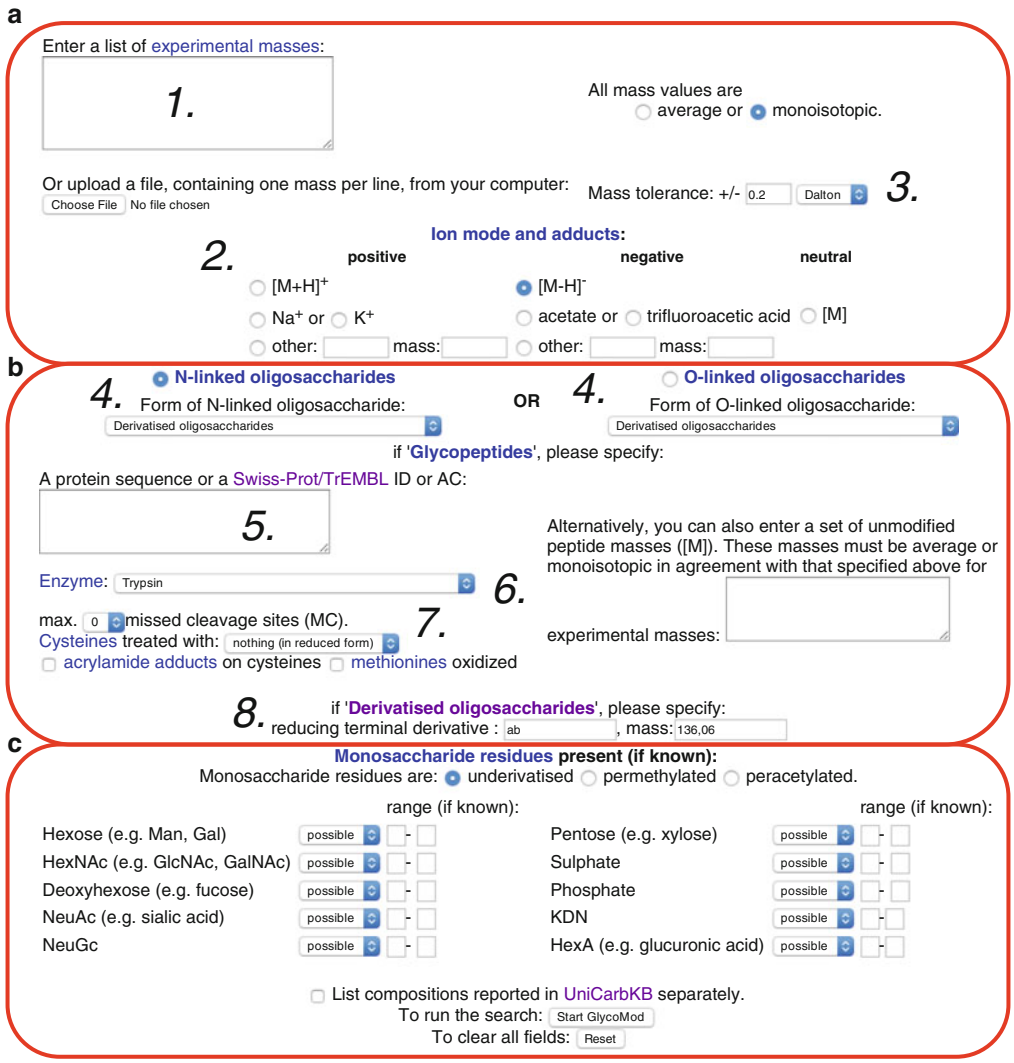


Fig. 2 Input page from Glycomod, showing: (**a**) area for input of MS parameters; (**b**) input of glycoconjugate metadata; and (**c**) monosaccharide compositional data. *Numbers* refer to the text in Subheading 3.1.1

Table 2 Results from *N*-linked test data using GlycoMod

Composition	Submitted [M – H]	Glycoform mass (Da)	Δmass (Da)
$(HexNAc)_2\ (Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$	1581.92	1444.534	0.318
$(Hex)_1\ (HexNAc)_2\ (Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$	1744.00	1606.587	0.345
$(Hex)_2\ (HexNAc)_2\ (Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$	1906.01	1768.64	0.302

Table 3 Results from *O*-linked test data using GlycoMod

Composition	Submitted [M – H]	Glycoform mass (Da)	Δmass (Da)
$(Hex)_1$ $(HexNAc)_1$	384.1	365.132	–0.05
$(Hex)_1$ $(HexNAc)_1$ $(NeuAc)_1$	675.1	656.228	–0.146
$(Hex)_1$ $(HexNAc)_1$ $(NeuAc)_2$	966.2	947.323	–0.141
$(Hex)_2$ $(HexNAc)_2$ $(NeuAc)_2$	1331.2	1312.455	–0.273

1. Start GlycoWorkbench and draw the putative glycan structure. Go to *Structure* > *Add structure*, and choose the type of glycan to draw on canvas. The structure can be further modified by other functional tools in *Structure* drop-down list, such as *Add residue* and *Add terminal*.
2. To set the theoretical mass, go to *Tools* > *Mass options of all structures*, where theoretical mass can be changed according to different settings.
3. To paste the MS/MS peak list to GlycoWorkbench (Fig. 3a), copy the list of fragments *m/z* and intensity from software such as Microsoft Excel, or directly copy from MS software. (*see* **Note 6**)
4. To calculate the theoretical fragmentation ions of the drawn structure, select *Tools* > *Fragments* > *Compute fragments for selected structures*. The window of *Fragment options* appears. After selecting glycosidic and cross-ring cleavage, click *OK* (Fig. 3b). The theoretical fragmentation list will appear in right panel (Fig. 3c).
5. To associate experimental MS/MS peak list with a computed peak list, go to *Tools* > *Annotation* > *Annotate peaks with fragments from selected structures*. A *Fragment option* window will appear. After setting the parameter, click *OK*. Then the associated peak list will appear in right panel (Fig. 3d).

3.3 From Analytical MS and MS/MS Data to Structure: UniCarb-DB

Different searches have been implemented for retrieving data from UniCarb-DB:

(a) Taxonomy/Tissue, where entries are selected based on their biological context,

(b) Mass and composition. As the name indicates, the search can be performed by mass and/or composition. Additional parameters can be entered to make the query more specific.

(c) MS/MS search, or peak-matching tool, retrieves and scores the peak lists in the database according to their similarity to a

Fig. 3 Annotation of experimental MS/MS peak list using GlycoWorkbench. (**a**) Input experimental MS/MS peak list into GlycoWorkbench; (**b**, **c**) Computing theoretic glycosidic and cross-ring cleavage using GlycoWorkbench; (**d**) Associating experimental MS/MS peak list with computed fragment of selected structure

target peak list. The minimum input data is a peak list in plain text separated by a tab stop character (Fig. 4a). These data can be copied and pasted directly into the search page from software such as Microsoft Excel, or directly copied from MS software.

3.3.1 Mass and Composition

1. Upon accessing UniCarb-DB on www.unicarb-db.org click on the "Get Started" button at the bottom of the page.
2. Select the option "Search" from the menu located at the top of the interface. Afterwards, click on the option "Mass and composition" from the menu located at the left side of the search page.
3. Locate the "composition" block at the bottom of the page. Glycan composition can be assembled by selecting the option "yes" from the respective drop-down list if the desired residue must be included in the structure. "No" is chosen if the residue must not be part of the structure. The default value "possible" indicates that the residue could be but does not have to be present in the displayed results. The text boxes following the lists are used to specify the number of residues (exact or interval) that are expected. For example, use the values "yes"

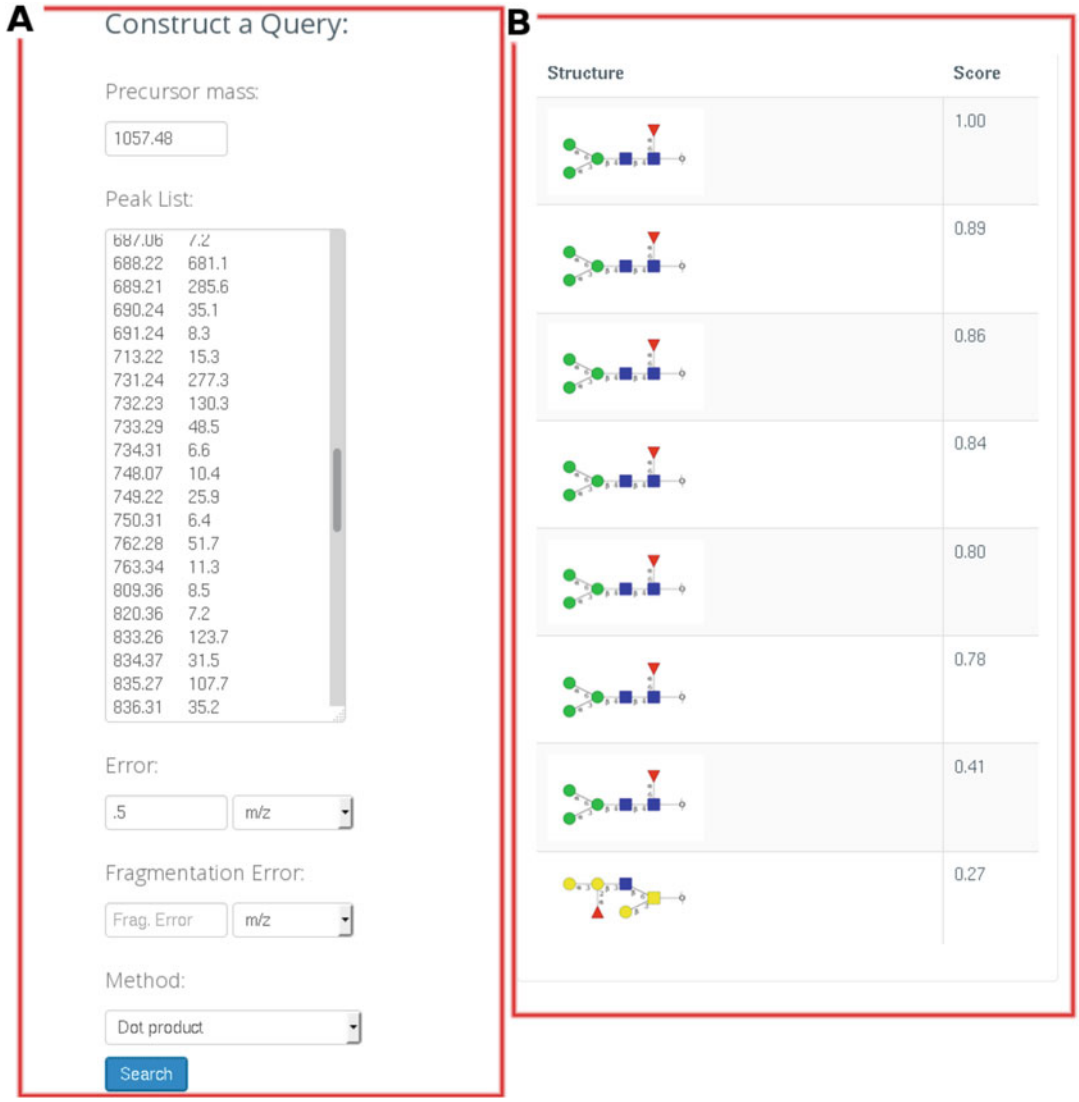


Fig. 4 UniCarb-DB MS/MS query. (**a**) Input page of the MS/MS search page on UniCarb-DB and (**b**) result page of submitted MS/MS fragment data using normalized dot product

and "1" for "dHex", "yes" and "1"-"2" for Hex, and "yes" and "2" for HexNAc.

4. Click on the "search" button.
5. The results show the list of structures that satisfy the constraints. The result can be refined using the rest of the parameters on the search form.
6. Return to the "Mass and composition" search form and perform all the previous steps.

7. Input a mass value. For example, "894.3" and select the option "Calculated".
8. The error parameter should be included only when the mass value is specified. The error is calculated for both upper and lower ranges. For example, if the value "0.5" is used, the query will select all spectra whose calculated mass values lie between 894.8 and 893.8. The drop-down list next to the text box includes three measurement formats for the error value: *m/z*, ppm, and dalton. Leave the option selected for default (*m/z*).
9. Select the option "–", "underivatised", and "reduced" from the drop-down lists "charge state", "derivative", and "reducing end", respectively.
10. Click on the "search" button.

3.3.2 Peak Matching Tool (MS/MS Search)

1. Upon accessing UniCarb-DB on www.unicarb-db.org click on the "Get Started" button at the bottom of the page.
2. Select the option "Search" from the menu located at the top of the UniCarb-DB interface. Afterwards, click on the option "MS/MS" from the menu located at the left side of the search page.
3. MS/MS spectra are represented as peak lists that consist of two columns separated by a tab stop character. The first column corresponds to the *m/z* values and the second to the respective intensities. Paste the target peak list from the excel file into the proper text box in the search form (Fig. 4a). This is the only requirement to start the search; however, the query can take a long time to display the results. The use of additional parameters can speed up the process notably.
4. To continue with the test case, use the value "1057.48" for the precursor mass.
5. The error parameter should be included only when the precursor mass is also being used. The error is calculated for both upper and lower ranges. For example, with an error value of "0.5", the query will select all spectra whose precursor mass values lie between 1056.98 and 1057.98. The drop-down list next to the text box includes three measurement formats for the error value: *m/z*, ppm, and Dalton. For the test case, leave the default option (*m/z*).
6. The fragmentation error is used to adjust for possible discrepancies among representations of numeric values.
7. The "method" drop-down list includes, for the time being, only the option "dot product". This indicates the type of algorithm that is used to perform the comparison.

8. The results are displayed as a list of structures, each one with a score (Fig. 4b). The higher the score, the more similar to the peak the lists are.

3.4 From NMR Data to Structure: CASPER

In the following example, an NMR chemical shift prediction and a structural elucidation of trisaccharide [33] and tetrasaccharide [34] *O*-mannosyl glycans present on the glycoprotein α-dystroglycan are presented using components. For structural determination, unassigned NMR spectra are also used as input. CASPER will, for a structural elucidation, produce a ranked list of structural suggestions based on the best fit of predicted chemical shifts to those from NMR experiments. Furthermore, tentative ^{1}H and ^{13}C NMR chemical shift assignments are made, as in the prediction mode, for all resonances; and a schematic chemical shift map (HSQC-type) that compares experimental and predicted chemical shifts is generated. In addition, a three-dimensional structure in pdb-format can be made seamlessly using the CarbBuilder program [35] from within the "CASPER results menu". A detailed step-by-step description of how CASPER is used was recently reported [36].

3.4.1 Input to CASPER for Prediction of NMR Chemical Shifts

In the present case, using the previously mentioned *O*-mannosyl glycan trisaccharide initiated by POMGNT2 [33], the input is prepared in the following way (Fig. 5a):

1. Go to the "Calculate Chemical Shifts" option in CASPER. (www.casper.organ.su.se/casper/calculate.php)
2. Enter a title and/or literature source for the project. (This is optional but will help with keeping track of multiple projects.) It is also possible to load a previous CCPN project [37].
3. Enter the constituents of the target structure. Anomeric configuration, residue, where the residue links, and to which residue it links must be specified. Several options for non-reducing sugars, such as O-Me, are also available. In the case of *N*-glycans and *O*-glycans, amino acids are specified as residues; however, the anomeric configuration does not apply. If it is a repeating unit, the "reducing end" should be linked to the terminal residue of the repeating unit of the backbone in the polymer. Note that the residues are specified in a biosynthetic/reducing end order, i.e., the threonine would be entered first in the example above, followed by the mannose residue. This is a truncated view, and up to 12 residues can be entered.
4. If NMR chemical shifts are known, they can be added. In the resulting output, the differences between the entered chemical shifts and the calculated ones are displayed. This is true for both ^{1}H and ^{13}C NMR data.

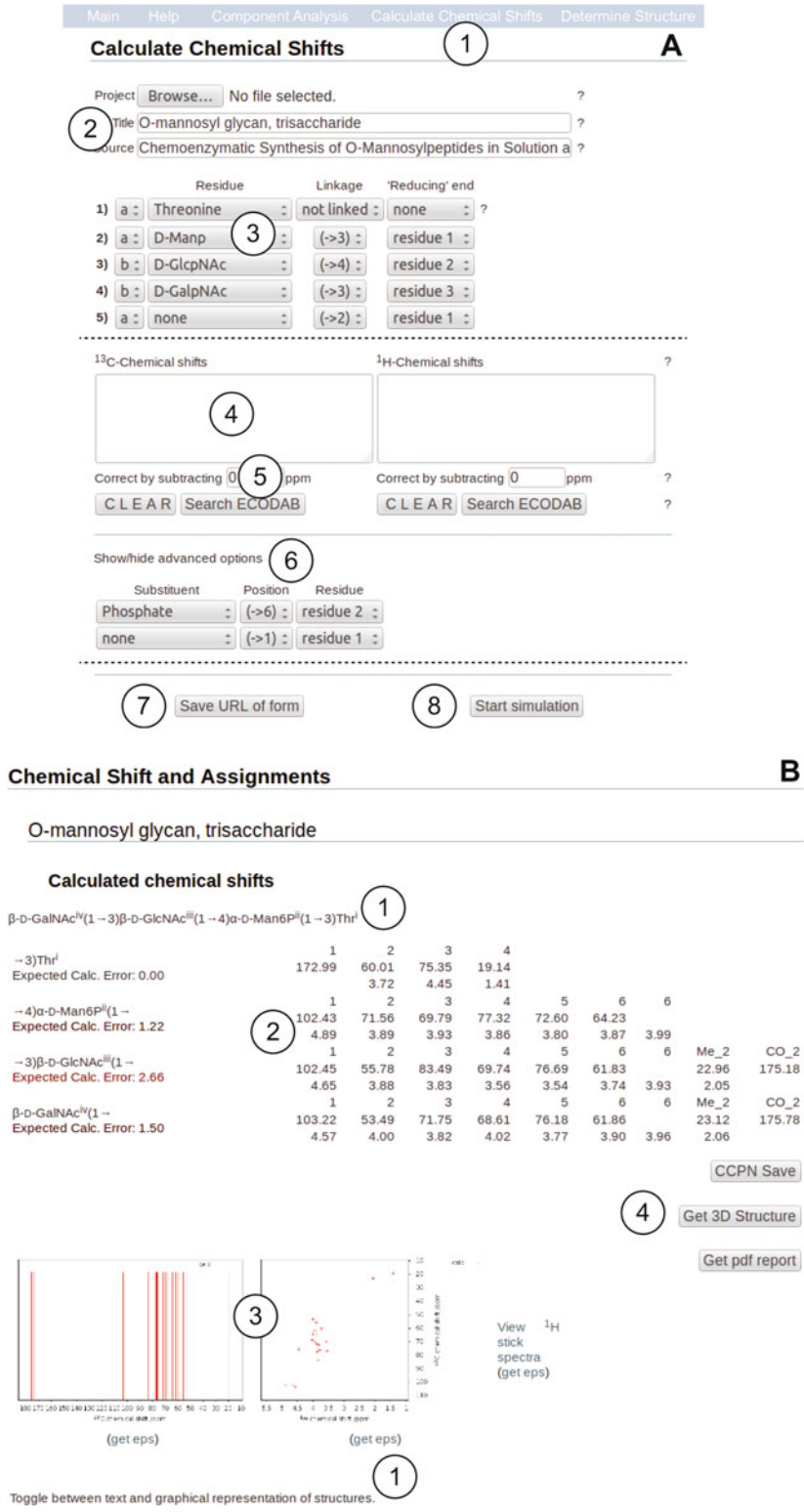


Fig. 5 CASPER input and output. (**a**) Example input for a CASPER NMR chemical shift prediction of an *O*-mannosyl glycan found on human α-dystroglycan. A *dashed line* indicates a truncated view of the part above it. (**b**) results for a CASPER NMR chemical shift prediction of an *O*-mannosyl glycan found on human α-dystroglycan. (**c**) example input for a CASPER structure determination of an *O*-mannosyl glycan found on human α-dystroglycan. The toggled window for 2D data input is not shown. A *dashed line* indicates a truncated view of the part above it

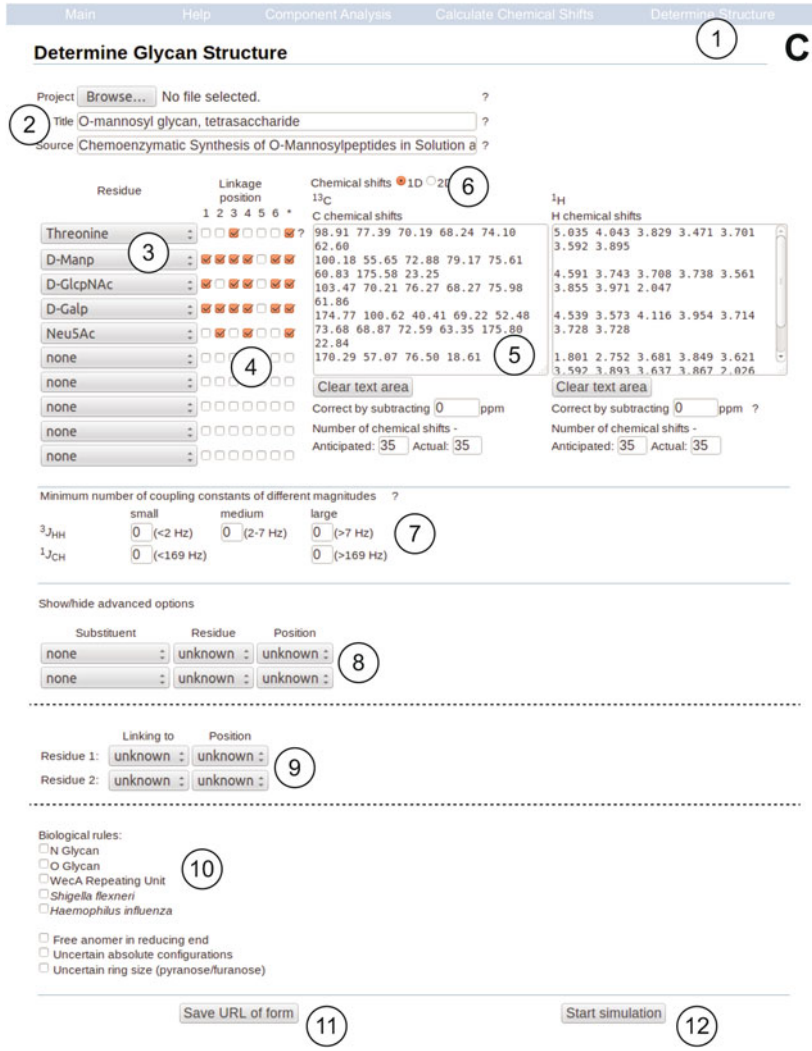


Fig. 5 (continued)

5. A systematic correction can also be added, for example, when NMR chemical shift data is obtained using different reference compounds.
6. The display for advanced options can be toggled; and, in the "Calculate Chemical Shifts" case, the option is to add substituents. This is done in a similar manner similar to specifying residues. Once again, a truncated view is shown, and up to 12 substituents can be entered.
7. The URL can be saved to quickly enter a previous input; this is done by clicking "Save URL of form".
8. Submit the job to CASPER by pressing "Start simulation". Once the calculation is complete, the page will be redirected to a results page (*see* **Note** 7).

3.4.2 Results Page from CASPER for Prediction of NMR Chemical Shifts

Once the calculation is complete, the user will automatically be redirected to a results page (Fig. 5b).

1. At the top, the structure used as input is shown in the CASPER-format [19]. The structure can also be toggled to display the CFG representation [38].
2. The NMR chemical shifts are displayed in a list for each residue along with an "expected calc. error". If chemical shift approximations to the structural entities have been made, a lower value indicates higher similarity to the predicted chemical shifts. This is a measure of the expected accuracy of the calculation.
3. Graphical representations of ^{13}C and $^{1}H/^{13}C$-correlated data are generated. If needed, the ^{1}H data can also be displayed in stick representation. It is also possible to download these in the Encapsulated PostScript format.
4. The results can be saved as a CCPN project or as a pdf-report. The option "Get 3D Structure" generates a 3D structure in the pdb format using a local version of CarbBuilder [35].

3.4.3 Using CASPER to Predict Structures from NMR Chemical Shifts

It is also possible to use CASPER to determine a structure (Fig. 5c) from NMR chemical shifts along with any known constituents, for example, from sugar analysis, methylation analysis, or putative glycosyltransferase functions retrieved from a gene sequence comparison [21, 22]. In this example, the structure of an *O*-mannosyl glycan tetrasaccharide is determined based on previously published data [34].

1. Go to the "Determine Structure" option in CASPER. (www.casper.organ.su.se/casper/determine.php)
2. As in the "Calculate Chemical Shifts" mode, a brief title and source can be entered to help keep track of projects. This is optional.
3. Enter the residues, known or unknown. Unknown residues will increase calculation times significantly. Since linkages are not specified here, the order of residues does not matter; however, the first drop-down window does have some additional options, so the residue expected to be at the reducing end must be entered here; in our case the threonine residue is given.
4. Enter the possible linkage positions. The numbering 1–6 are all AND statements, but use of the wildcard (*) will allow the linkage positions to be AND/OR statements.
5. Enter the ^{13}C and/or ^{1}H NMR chemical shifts. Just as in predicting chemical shifts, if there is a systematic error, a correction to the chemical shifts can be added.

6. By toggling to 2D experiments, data from HSQC/HETCOR, TOCSY, HMBC, and/or H2BC experiments can also be supplied (toggled window not shown).
7. If the coupling constants ($^3J_{HH}$, $^1J_{CH}$) for the anomeric protons are known, they can be entered to limit the number of structures for which NMR chemical shift calculations are carried out, thereby speeding up the procedure.
8. Advanced options can be toggled to specify any substituent(s) present; and if known, where the residues are attached and in which position. This is a truncated view; up to ten substituents can be entered.
9. There is also an option to specify linkages. This will override any ambiguity given in step C4. All ten residues can have linkage positions specified.
10. Finally, there are several options to make the search more restricted, e.g., the allowed structural elements can be limited according to biological rules, or they can be permissive. For example, when the anomeric configuration at the reducing end or the absolute configurations are unknown, all options are considered.
11. The URL can be saved to quickly enter a previous input; this is done by clicking "Save URL of form".
12. Submit the job to CASPER by pressing "Start simulation". Once the calculation is complete, the page will be redirected to a results page (*see* **Note 8**).

3.5 From HPLC/UHPLC/CE Data to N-linked Structure: GlycoBase

GlycoBase (Fig. 6) contains several glycan collections; all searches can be performed on the entire collection, a selected subsection or on glycans derived from a particular sample within a collection. One can apply different search parameters, including features such as name, antennary composition (e.g., A2), GU value, particular glycan feature (e.g., sialic acid), stoichiometry (e.g., number of hexoses), mass, or biological context (e.g., tissue name or disease.)

3.5.1 Preliminary Annotations Based on Undigested Profile

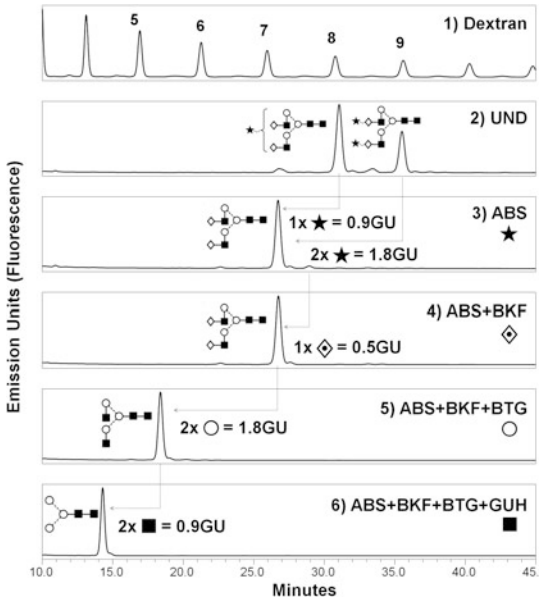
1. Launch GlycoBase
2. As a first approach, a search of the entire collection is recommended to ensure complete coverage of the glycan spectrum.
3. You can arrange all glycan structures based on increasing or decreasing GU value by simply clicking "HILIC-HPLC".
4. Locate the "Simple Search" field and search for a GU value that you have previously obtained, or one from the example data set.
5. The first glycan peak in the example data (Fig. 7, Table in Figure) has a GU value of 7.15. By searching the database, you will discover that one potential glycan of the same GU value is the A2(4)G2 structure.

Fig. 6 An overview of GlycoBase components: (**a**) GlycoBase's homepage with a simple and intuitive interface; (**b**) Search tool that allows you to search for glycan structures in a number of ways; (**c**) a table obtained from an example search; (**d**) the description and summary for A2G2 structure. Each glycan entry contains information about properties, isomers, average GU values from different modes of separation, an option to display a glycan cartoon in different notation, and a link to view an entry in UniCarb-DB; (**e**) exoglycosidase digestion data including published digestions, and records of internal digestions

6. Continue searching the database for the next peak with a GU value of 8.06. The glycan structure with the most similar GU value is FA2G(4)2S(6)1; however, it is advisable to make a note of other potential glycans with similar GU values.
7. Continue searching the database for remaining glycan peaks, and write down all possibilities.
8. For a single HPLC peak, there could be several options with potential structures of similar GU value. Preliminary annotations obtained from the GlycoBase search can be verified with exoglycosidase digestions.

3.5.2 Confirmation of Annotations with Exoglycosidase Digestions

1. Glycan sequencing by digestion with exoglycosidases is a very powerful tool for elucidating glycan structures. The enzymes may be used in a single or panel digestion array. They cleave specific glycosidic bonds between individual monosaccharides



1_Fibrinogen_UND	% Area	GU	Structure
Peak 1	4.42	7.15	A2G2
Peak 2	55.06	8.06	A2G2S1
Peak 3	1.44	8.26	A2BG2S1
Peak 4	4.18	8.56	FA2G2S1
Peak 5	33.65	9.00	A2G2S2
Peak 6	0.73	9.19	A2BG2S2
Peak 7	0.51	9.41	FA2G2S2
2_Fibrinogen_ABS	**% Area**	**GU**	
Peak 1	94.09	7.15	A2G2
Peak 2	2.43	7.32	A2BG2
Peak 3	3.48	7.61	FA2G2
3_Fibrinogen_ABS+BKF	**% Area**	**GU**	
Peak 1	97.09	7.15	A2G2
Peak 2	2.91	7.32	A2BG2
4_Fibrinogen_ABS+BKF+BTG	**% Area**	**GU**	
Peak 1	97.86	5.36	A2
Peak 2	2.14	5.51	A2B
5_Fibrinogen_ABS+BKF+BTG+GUH	**% Area**	**GU**	
Peak 1	97.74	4.32	M3
Peak 2	2.26	4.45	M3B

Fig. 7 Example data acquired from 2AB-labeled glycan separation from fibrinogen containing (1) dextran ladder with indicated GU values, (2) undigested profile of fibrinogen with major peaks shown, (3) sialidase digested, (4) sialidase and fucosidase digested, (5) sialidase, fucosidase, and galactosidase digested, (6) sialidase, fucosidase, galactosidase, *N*-acetylhexosaminidase digested. *Symbols* denoted on the *right* of each chromatogram indicate the individual monosaccharides removed through exoglycosidase digestion. The *table* on the *right* contains detected peaks with assigned GU values and relative % areas for all peaks, as well as their annotated structures

units from the non-reducing end, and are highly specific for both the terminal residue and its linkage type.

2. An example of an exoglycosidase digestion array from fibrinogen is shown in Fig. 7.
3. Confident assignments of structures can be made by comparing chromatographic profiles obtained before and after digestion with exoglycosidases of known specificity. Apply the strategy of top-down digestion and bottom-up identification.
4. Start with the first digestion: ABS (*Arthrobacter ureafaciens* sialidase) releases α2-3, α2-6, and α2-8 linked non-reducing terminal sialic acid residues. Observe the peaks that are eliminated with sialidase treatment. Peaks that are removed are susceptible to sialidase treatment, which implies that they contain a sialic acid residue at the terminal end. To further validate linkage type, additional digestions with an NANI sialidase may release α2-3 linked non-reducing terminal sialic acid.
5. Continue evaluating the subsequent digestions and observe shifts in peak positions that correspond to particular glycan residues: BKF (bovine kidney α-fucosidase) releases α1-2,3,4 and 6-linked non-reducing terminal fucose residues.

BTG (bovine testes β-galactosidase) hydrolyses non-reducing terminal β1-3- and β1-4-linked galactose, and GUH (hexosaminidase cloned from *Streptococcus pneumoniae* expressed in *E. coli*) releases GlcNAc residues linked to Man (Fig. 7).

6. Confirm potentially identified structures by checking their GU value with the database.
7. Always double check the exported table containing GU values and the corresponding relative percentage areas. Area values of species before and after a digest should match.
8. In the case of complex samples it is possible to collect individual peaks. The volatile buffers can be removed and the glycans can be analyzed by mass spectrometry to validate assignments. The combination of different orthogonal techniques increases the accuracy of structure identification.

3.6 Validation of Monosaccharide Sequence and Linkages with the Aid of Exoglycosidase Digestion: GlycoDigest

A description of the tool including example data and details of exoglycosidases can be found at www.glycodigest.org. From the homepage, links guide users to the "Getting Started Guide" hosted on GitHub at bit.ly/1lZlcFk, and to "Exoglycosidase Information" options and example data. The web application version of GlycoDigest has been integrated into UniCarbKB.

3.6.1 Step One: Build Glycan Structure

1. Launch UniCarbKB-GlycoDigest (unicarbkb.org/builderDigest) (*see* **Note 9**).
2. The latest version of GlycanBuilder offers the same functionality as that developed by the EUROCarbDB project. To start building a structure, choose one of the following three options: (1) select monosaccharides from the canvas panel or "Structure" menu; (2) extend a structure from a predefined list; or (3) build a substructure/epitope.
3. Structures can be built by adding new monosaccharides or predefined groups to the highlighted residue. New groups can be selected either from the medium toolbar or from the structures menu. The list of structural constituents comprises an exhaustive collection of saccharides, substituents, reducing-end markers, and saccharide modifications. All stereochemical information about a saccharide, such as anomeric conformation, chirality, ring configuration, and linkage position, can be specified.
4. After building the appropriate glycan structure, select the "Digest" button.

3.6.2 Step Two: Selecting Exoglycosidases

The next step is to theoretically digest the structure with a defined array of exoglycosidases. Careful design and formulation of exoglycosidase arrays is critical to the success of correct glycan structural assignments.

1. To create an array of exoglycosidases, use the "Build Exoglycosidase Array" box to select supported exoglycosidases. (*see* **Note 10**)
2. After selecting a single endoglycosidase or panel of enzymes, select the "Digest" button.

3.6.3 Step Three: Output

GlycoDigest will simulate the combinations of exoglycosidase selected in Subheading 3.6.2. When complete, the "Results" section will be updated with theoretically computed end-product(s).

3.7 Association of Structure with Biology: UniCarbKB and Association of Glycan Structures with Proteins

UniCarbKB is an initiative that provides an online information storage and search platform for glycomics and glycobiology research. The knowledgebase offers open access to a rich resource supported by querying interfaces, annotation technologies, and the adoption of common standards to integrate structural, experimental, and functional data.

To centralize and organize the features of UniCarbKB, a confluence knowledge base is available (confluence.unicarbkb.org). The knowledgebase is a repository of how-to and troubleshooting guides that provides users with access to solutions and relevant search results.

The confluence knowledge base is split into four main sections: (1) Content, (2) Browsing, (3) Nomenclature, and (4) How-to articles. Each section is further divided into specific layers that focus on a particular feature, which includes screenshots and step-by-step notes on discovering data sets and navigating the database.

1. Launch UniCarbKB Confluence (confluence.unicarbkb.org)
2. Select a section from the navigation panel on the left-hand side. For example, to find information about glycoproteins click the "How-to articles" link followed by the "Glycoprotein Page".
3. Each page summarizes the content available with documented screenshots and how users can access relevant information. For the "Glycoprotein Page" a screenshot highlights the major content sections and briefly describes the level of information available.
4. The "Related articles" section provides links to related guides.
5. The "Search" box offers a quick navigation aid as well as a full site search. When you perform a full search, Confluence will search all content and the results screen shows a list of matching items.

The "How-to articles" provide new and returning users with access to up-to-date documented guides. For example, the "Reference" guide provides an overview for selecting and searching references and a screenshot of an individual reference page.

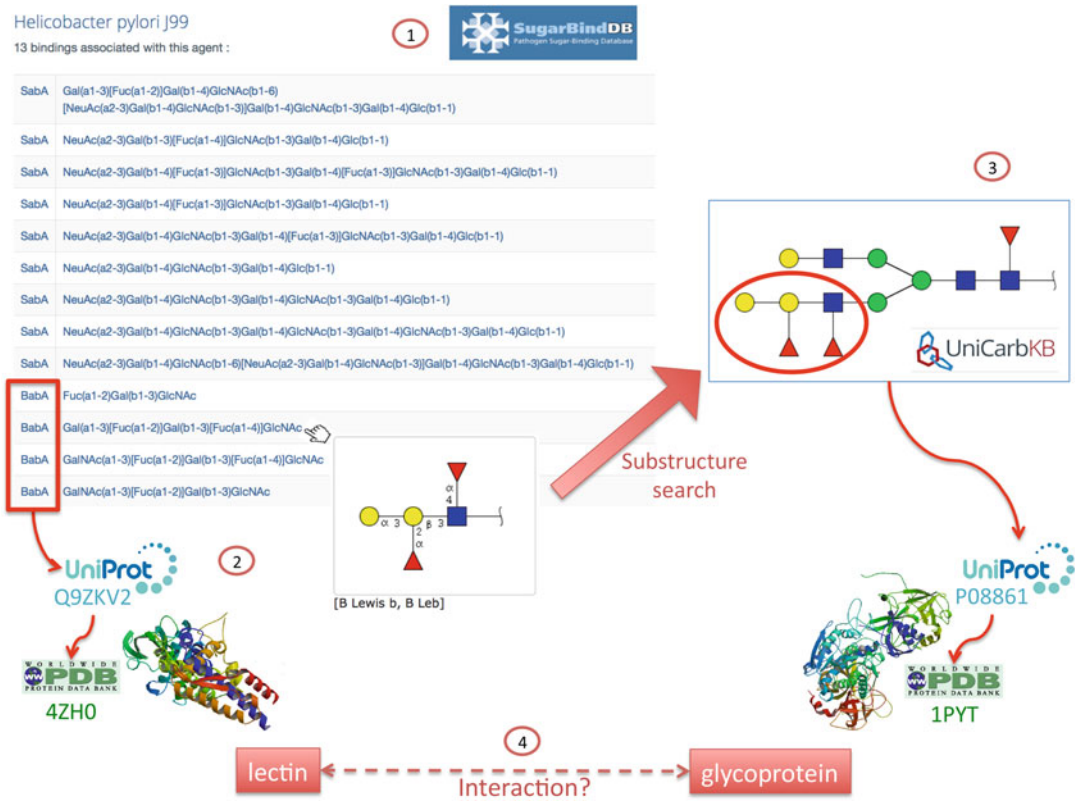


Fig. 8 Investigating glycan-mediated protein–protein interactions using SugarBind. (1) shows the output list of protein–glycan bindings when SugarBindDB is queried with "Helicobacter pylori J99". In this list, the BabA-B Lewis b (i.e., Galα1-3(Fucα1-2)Galβ1-3(Fucα1-4)GlcNAc) is selected. Mousing over the ligand written in IUPAC code prompts an image of the B Lewis b written in CFG symbols, as shown in (2). The BabA lectin is cross-referenced to UniProt entry Q9ZKV2. (3) shows the result of searching UniCarbKB with B Lewis b as a potential substructure of full glycan structures. Only one match is found. This UniCarbKB entry is linked to a single UniProt entry: P08861. Finally the lectin and glycoprotein information can be correlated via glycans and (4) highlights the potential for interaction

3.8 Association of Structure with Microbial Interactions: SugarBind

A glycan-lectin binding pair within SugarBindDB can lead to the identification of a glycan-mediated protein–protein interaction, i.e., a lectin–glycoprotein interaction. The following example describes how this information can be revealed.

1. Search SugarBindDB for a specific pathogen, for instance *Helicobacter pylori* J99. A list of glycan-binding proteins (BabA and SabA lectins) is displayed along with all stored glycan ligands.
2. In the list of output glycan-binding pairs, consider the case of BabA-B Lewis b (i.e., Galα1-3(Fucα1-2)Galβ1-3(Fucα1-4)GlcNAc) (Fig. 8). Click on the glycan link, which prompts a new page that displays the structural details of B Lewis b.

3. From the B Lewis b page, follow the cross-link to UniCarbKB via the substructure search, which reveals the structure of a corresponding full glycan entry (UniCarbKB 7563) that contains B Lewis b. According to UniCarbKB information, this glycan is reported to be attached to Asn-114 of human elastase 1 (P08861).
4. Hypothesize the potential interaction between *H. pylori* BabA lectin (Q9ZKV2) with the human glycoprotein elastase 1 (P08861).

Needless to say, this result is hypothetical and requires confirmation. However, it provides a lead for further investigation of correlations between Lewis B epitope expression and the severity of diseases caused by *Helicobacter pylori* strains expressing BabA adhesins [39].

4 Notes

1. The GlycoMod tool is specifically designed for *N*-linked and *O*-linked glycoproteins. The possible allowed *N*-linked compositions are more restricted by the presence of the core pentasaccharide (GlcNAc2, Man3) than the *O*-linked; so, for generic searches of other types of glycoconjugates, the *O*-linked search is recommended.
2. Since the database currently consists mainly of negative ion mode MS/MS fragment spectra, other types of data are welcome, and should be submitted as glycoworkbench files.
3. Charge deconvolution of MS data is required if the charge state of pseudomolecular ions is higher than ±1.
4. For mammalian sugars, monosaccharides in the left hand column (7) should be left as selection "possible", while the right hand column should be selected only if one has additional knowledge or indication that these sugars are present in the sample. (for example, humans usually do not express NeuGc.)
5. The term "glycoform mass" refers to an uncharged mass of a glycan without its reducing end modification, i.e., it should be considered to be the mass of the reducing glycan after subtracting the mass of a water molecule.
6. For computer operating systems that use "comma (,)" as decimal point, all occurrences of comma should be changed to "dot (.)" before pasting a MS/MS peak list to GlycoWorkbench. For Xcalibur, one can export the list directly using *Export> Clipboard* (*Exact Mass*) to GlycoWorkbench. When pasting a peak list, Crtl + V does not work.
7. An example input can be accessed from the following link: http://tinyurl.com/CalcShiftTri.

8. An example input for the tetrasaccharide above can be accessed from the following link: http://tinyurl.com/DetermineTetra.
9. To successfully model the action of an exoglycosidase, the monosaccharide anomeric form and glycosidic linkages must be defined. Exoglycosidases are highly specific, and the rules embedded in GlycoDigest will only work if the submitted structure is fully defined. To specify linkage and anomeric form, select a residue and configure the residue using the main GlycanBuilder canvas.
10. Abbreviated names of exoglycosidases are used, but a pop-up table is available that summarizes their specificity. To access the table, select the "Information on Exoglycosidases" button.

Acknowledgements

The authors acknowledge support from the European Union FP7 GastricGlycoExplorer ITN under grant agreement no.316929, Petrus and Augusta Hedlunds foundation, The Swedish Foundation for International Cooperation in Research and Higher Education, and the Swedish Research Council (621-2013-5895 and 621-2013-4859).

References

1. Doubet S, Bock K, Smith D, Darvill A, Albersheim P (1989) The complex carbohydrate structure database. Trends Biochem Sciences 14(12):475–477
2. von der Lieth CW, Freire AA, Blank D, Campbell MP, Ceroni A, Damerell DR, Dell A, Dwek RA, Ernst B, Fogh R, Frank M, Geyer H, Geyer R, Harrison MJ, Henrick K, Herget S, Hull WE, Ionides J, Joshi HJ, Kamerling JP, Leeflang BR, Lütteke T, Lundborg M, Maass K, Merry A, Ranzinger R, Rosen J, Royle L, Rudd PM, Schloissnig S, Stenutz R, Vranken WF, Widmalm G, Haslam SM (2011) EUROCarbDB: an open-access platform for glycoinformatics. Glycobiology 21(4):493–502. doi:10.1093/glycob/cwq188
3. Cooper CA, Harrison MJ, Wilkins MR, Packer NH (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. Nucleic Acids Res 29(1):332–335
4. Hayes CA, Karlsson NG, Struwe WB, Lisacek F, Rudd PM, Packer NH, Campbell MP (2011) UniCarb-DB: a database resource for glycomic discovery. Bioinformatics 27(9): 1343–1344. doi:10.1093/bioinformatics/btr137, btr137 [pii]
5. Campbell MP, Hayes CA, Struwe WB, Wilkins MR, Aoki-Kinoshita KF, Harvey DJ, Rudd PM, Kolarich D, Lisacek F, Karlsson NG, Packer NH (2011) UniCarbKB: putting the pieces together for glycomics research. Proteomics 11(21):4117–4121. doi:10.1002/pmic.201100302
6. Ranzinger R, Herget S, Wetter T, von der Lieth CW (2008) GlycomeDB - integration of open-access carbohydrate structure databases. BMC Bioinformatics 9:384. doi:10.1186/1471-2105-9-384
7. Royle L, Campbell MP, Radcliffe CM, White DM, Harvey DJ, Abrahams JL, Kim YG, Henry GW, Shadick NA, Weinblatt ME, Lee DM, Rudd PM, Dwek RA (2008) HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. Anal Biochem 376(1):1–12. doi:10.1016/j.ab.2007.12.012, S0003-2697(07)00794-4 [pii]
8. Hayes CA, Carta G, Karlsson NG, Duffy F, Rudd PM (2015) Informatics and Analytical Tools for Glycan Analysis and the Development of Biot. In: Cipolla L (ed) Carbohydrate chemistry: state of the art and challenges for drug development: an overview on structure,

biological roles. Synthetic methods and application as therapeutics. Imperial College Press, London, pp 173–192

9. Campbell MP, Nguyen-Khuong T, Hayes CA, Flowers SA, Alagesan K, Kolarich D, Packer NH, Karlsson NG (2014) Validation of the curation pipeline of UniCarb-DB: building a global glycan reference MS/MS repository. Biochim Biophys Acta 184(1 Pt A):108–116. doi:10.1016/j.bbapap.2013.04.018
10. Cooper CA, Gasteiger E, Packer NH (2001) GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. Proteomics 1(2):340–349. doi:10.1002/1615-9861(200102)1:2<340::AID-PROT340>3.0.CO;2-B
11. Ceroni A, Maass K, Geyer H, Geyer R, Dell A, Haslam SM (2008) GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. J Proteome Res 7(4):1650–1659. doi:10.1021/pr7008252
12. Damerell D, Ceroni A, Maass K, Ranzinger R, Dell A, Haslam SM (2012) The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. Biol Chem 393(11):1357–1362. doi:10.1515/hsz-2012-0135
13. Jansson P-E, Kenne L, Widmalm G (1987) CASPER—a computerised approach to structure determination of polysaccharides using information from n.m.r. spectroscopy and simple chemical analyses. Carbohydr Res 168 (1): 67-77.doi:http://dx.doi.org/10.1016/0008-6215(87)80007-1
14. Jansson PE, Kenne L, Widmalm G (1989) Computer-assisted structural analysis of polysaccharides with an extended version of CASPER using ^{1}H- and ^{13}C-n.m.r. data. Carbohydr Res 188:169–191
15. Jansson PE, Kenne L, Widmalm G (1991) CASPER: a computer program used for structural analysis of carbohydrates. J Chem Inf Comput Sci 31(4):508–516
16. Stenutz R, Jansson PE, Widmalm G (1998) Computer-assisted structural analysis of oligo- and polysaccharides: an extension of CASPER to multibranched structures. Carbohydr Res 306(1-2):11–17
17. Jansson PE, Kenne L, Widmalm G (1991) Computer-assisted structural analysis of oligosaccharides using CASPER. Anal Biochem 199(1):11–17
18. Jansson PE, Stenutz R, Widmalm G (2006) Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel Web-based version of the computer program CASPER. Carbohydr Res 341(8):1003–1010. doi:10.1016/j.carres.2006.02.034
19. Lundborg M, Widmalm G (2011) Structural analysis of glycans by NMR chemical shift prediction. Anal Chem 83(5):1514–1517. doi:10.1021/ac1032534
20. Lundborg M, Fontana C, Widmalm G (2011) Automatic structure determination of regular polysaccharides based solely on NMR spectroscopy. Biomacromolecules 12(11):3851–3855. doi:10.1021/bm201169y
21. Lundborg M, Modhukur V, Widmalm G (2010) Glycosyltransferase functions of E. coli O-antigens. Glycobiology 20(3):366–368. doi:10.1093/glycob/cwp185
22. Rojas-Macias MA, Ståhle J, Lütteke T, Widmalm G (2015) Development of the ECODAB into a relational database for Escherichia coli O-antigens and other bacterial polysaccharides. Glycobiology 25(3):341–347. doi:10.1093/glycob/cwu116
23. Roslund MU, Säwén E, Landström J, Rönnols J, Jonsson KH, Lundborg M, Svensson MV, Widmalm G (2011) Complete ^{1}H and ^{13}C NMR chemical shift assignments of mono-, di-, and trisaccharides as basis for NMR chemical shift predictions of polysaccharides using the computer program CASPER. Carbohydr Res 346(11):1311–1319. doi:10.1016/j.carres.2011.04.033
24. Ronnols J, Pendrill R, Fontana C, Hamark C, Angles d'Ortoli T, Engstrom O, Ståhle J, Zaccheus MV, Säwén E, Hahn LE, Iqbal S, Widmalm G (2013) Complete ^{1}H and ^{13}C NMR chemical shift assignments of mono- to tetrasaccharides as basis for NMR chemical shift predictions of oligosaccharides using the computer program CASPER. Carbohydr Res 380: 156–166. doi:10.1016/j.carres.2013.06.026
25. Royle L, Radcliffe CM, Dwek RA, Rudd PM (2006) Detailed structural analysis of N-glycans released from glycoproteins in SDS-PAGE gel bands using HPLC combined with exoglycosidase array digestions. In: Brockhausen I (ed) Meth Mol Biol, vol 347 Glycobiology Protocols. Humana Press Inc., Totowa, NJ, pp 125–143
26. Adamczyk B, Struwe WB, Ercan A, Nigrovic PA, Rudd PM (2012) Characterization of fibrinogen glycosylation and its importance for serum/plasma N-glycome analysis. J Proteome Res 12(1):444–454
27. Campbell MP, Royle L, Rudd PM (2015) GlycoBase and autoGU: resources for interpreting HPLC-glycan data. Methods Mol Biol 1273:17–28. doi:10.1007/978-1-4939-2343-4_2

28. Gotz L, Abrahams JL, Mariethoz J, Rudd PM, Karlsson NG, Packer NH, Campbell MP, Lisacek F (2014) GlycoDigest: a tool for the targeted use of exoglycosidase digestions in glycan structure determination. Bioinformatics 30(21):3131–3133. doi:10.1093/bioinformatics/btu425
29. Campbell MP, Peterson R, Mariethoz J, Gasteiger E, Akune Y, Aoki-Kinoshita KF, Lisacek F, Packer NH (2014) UniCarbKB: building a knowledge platform for glycoproteomics. Nucleic Acids Res 42(Database issue): D215–D221. doi:10.1093/nar/gkt1128
30. UniProt C (2015) UniProt: a hub for protein information. Nucleic Acids Res 43(Database issue):D204–D212. doi:10.1093/nar/gku989
31. Gaudet P, Michel PA, Zahn-Zabal M, Cusin I, Duek PD, Evalet O, Gateau A, Gleizes A, Pereira M, Teixeira D, Zhang Y, Lane L, Bairoch A (2015) The neXtProt knowledgebase on human proteins: current status. Nucleic Acids Res 43(Database issue):D764–D770. doi:10.1093/nar/gku1178
32. Mariethoz J, Khatib K, Alocci D, Campbell MP, Karlsson NG, Packer NH, Mullen EH, Lisacek F (2016) SugarBindDB, a resource of glycan-mediated host-pathogen interactions. Nucleic Acids Res 44(D1):D1243–D1250. doi:10.1093/nar/gkv1247
33. Live D, Wells L, Boons GJ (2013) Dissecting the molecular basis of the role of the O-mannosylation pathway in disease: α-dystroglycan and forms of muscular dystrophy. Chembiochem 14(18):2392–2402. doi:10.1002/cbic.201300417
34. Sardzik R, Green AP, Laurent N, Both P, Fontana C, Voglmeir J, Weissenborn MJ, Haddoub R, Grassi P, Haslam SM, Widmalm G, Flitsch SL (2012) Chemoenzymatic synthesis of O-mannosylpeptides in solution and on solid phase. J Am Chem Soc 134(10):4521–4524. doi:10.1021/ja211861m
35. Kuttel M, Yue M, Widmalm G, Lundborg M (2011) CarbBuilder: an adjustable tool for building 3D molecular structures of carbohydrates for molecular simulation. In: E-Science (e-Science), 2011 IEEE 7th international conference on, 5–8 Dec 2011. pp 395–402. doi:10.1109/eScience.2011.61
36. Lundborg M, Widmalm G (2015) NMR chemical shift prediction of glycans: application of the computer program CASPER in structural analysis. Methods Mol Biol 1273:29–40. doi:10.1007/978-1-4939-2343-4_3
37. Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59(4):687–696. doi:10.1002/prot.20449
38. Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Marth JD, Bertozzi CR, Hart GW, Etzler ME (2009) Symbol nomenclature for glycan representation. Proteomics 9(24):5398–5399. doi:10.1002/pmic.200900708
39. Gerhard M, Lehn N, Neumayer N, Boren T, Rad R, Schepp W, Miehlke S, Classen M, Prinz C (1999) Clinical relevance of the Helicobacter pylori gene for blood-group antigen-binding adhesin. Proc Natl Acad Sci U S A 96(22): 12778–12783
40. Farriol-Mathis N, Garavelli JS, Boeckmann B, Duvaud S, Gasteiger E, Gateau A, Veuthey AL, Bairoch A (2004) Annotation of post-translational modifications in the Swiss-Prot knowledge base. Proteomics 4(6):1537–1550. doi:10.1002/pmic.200300764
41. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
42. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 41(Database issue):D991–D995. doi:10.1093/nar/gks1193
43. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Ponten F (2015) Proteomics. Tissue-based map of the human proteome. Science 347 (6220): 1260419. doi:10.1126/science.1260419
44. Watanabe K, Yasugi E, Oshima M (2000) How to search the glycolipid data in LIPIDBANK for Web; the newly developed lipid database in Japan. Trends Glycosci Glyc 12(65):175–184. doi:10.4052/tigg.12.175
45. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH Jr, Murphy RC, Raetz CR, Russell DW, Subramaniam S (2007) LMSD: LIPID MAPS structure database. Nucleic Acids Res 35(Database issue):D527–D532. doi:10.1093/nar/gkl838
46. Agravat SB, Saltz JH, Cummings RD, Smith DF (2014) GlycoPattern: a web platform for glycan array mining. Bioinformatics

30(23):3417–3418. doi:10.1093/bioinformatics/btu559

47. Aoki-Kinoshita K, Agravat S, Aoki NP, Arpinar S, Cummings RD, Fujita A, Fujita N, Hart GM, Haslam SM, Kawasaki T, Matsubara M, Moreman KW, Okuda S, Pierce M, Ranzinger R, Shikanai T, Shinmachi D, Solovieva E, Suzuki Y, Tsuchiya S, Yamada I, York WS, Zaia J, Narimatsu H (2016) GlyTouCan 1.0—the international glycan structure repository. Nucleic Acids Res 44(D1):D1237–D1242. doi:10.1093/nar/gkv1041
48. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44(D1):D457–D462. doi:10.1093/nar/gkv1070
49. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M (2006) KEGG as a glycome informatics resource. Glycobiology 16(5):63R–70R. doi:10.1093/glycob/cwj010
50. Maeda M, Fujita N, Suzuki Y, Sawaki H, Shikanai T, Narimatsu H (2015) JCGGDB: Japan consortium for glycobiology and glycotechnology database. Methods Mol Biol1273:161–179.doi:10.1007/978-1-4939-2343-4_12
51. Tomiya N, Awaya J, Kurono M, Endo S, Arata Y, Takahashi N (1988) Analyses of N-linked oligosaccharides using a two-dimensional mapping technique. Anal Biochem 171(1):73–90
52. Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, Gupta R, Bennett EP, Mandel U, Brunak S, Wandall HH, Levery SB, Clausen H (2013) Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J 32(10):1478–1488. doi:10.1038/emboj.2013.79
53. Lütteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, von der Lieth CW (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. Glycobiology 16(5):71R–81R. doi:10.1093/glycob/cwj049
54. Perez S, Sarkar A, Rivet A, Breton C, Imberty A (2015) Glyco3D: a portal for structural glycosciences. Methods Mol Biol 1273:241–258. doi:10.1007/978-1-4939-2343-4_18
55. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42(Database issue):D490–D495. doi:10.1093/nar/gkt1178
56. Zhang H, Loriaux P, Eng J, Campbell D, Keller A, Moss P, Bonneau R, Zhang N, Zhou Y, Wollscheid B, Cooke K, Yi EC, Lee H, Peskind ER, Zhang J, Smith RD, Aebersold R (2006) UniPep—a database for human N-linked glycosites: a resource for biomarker discovery. Genome Biol 7(8):R73. doi:10.1186/gb-2006-7-8-R73
57. Kirschner KN, Yongye AB, Tschampel SM, Gonzalez-Outeirino J, Daniels CR, Foley BL, Woods RJ (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates. J Comput Chem 29(4):622–655. doi:10.1002/jcc.20820
58. Kunduru BR, Nair SA, Rathinavelan T (2016) EK3D: an E. coli K antigen 3-dimensional structure database. Nucleic Acids Res 44(D1):D675–D681. doi:10.1093/nar/gkv1313