

## Bioinformatics and molecular modeling in glycobiology

Martin Frank · Siegfried Schloissnig

Received: 23 December 2009 / Revised: 8 March 2010 / Accepted: 11 March 2010 / Published online: 4 April 2010  
© The Author(s) 2010. This article is published with open access at Springerlink.com

**Abstract** The field of glycobiology is concerned with the study of the structures, properties, and biological functions of the family of biomolecules called carbohydrates. Bioinformatics for glycobiology is a particularly challenging field, because carbohydrates exhibit a high structural diversity and their chains are often branched. Significant improvements in experimental analytical methods over recent years have led to a tremendous increase in the amount of carbohydrate structure data generated. Consequently, the availability of databases and tools to store, retrieve and analyze these data in an efficient way is of fundamental importance to progress in glycobiology. In this review, the various graphical representations and sequence formats of carbohydrates are introduced, and an overview of newly developed databases, the latest developments in sequence alignment and data mining, and tools to support experimental glycan analysis are presented. Finally, the field of structural glycoinformatics and molecular modeling of carbohydrates, glycoproteins, and protein–carbohydrate interaction are reviewed.

**Keywords** Glyco-bioinformatics · Databases · Carbohydrates · Glycosylation · Glycoproteins · Molecular modeling · Molecular dynamics simulation

### Introduction

The field of glycobiology is concerned with the study of the structures, properties, and biological functions of the family of biomolecules called carbohydrates. These carbohydrates can differ significantly in size ranging from monosaccharides to polysaccharides consisting of many thousands of carbohydrate units. One of the most significant features of carbohydrates is their ability to form branched molecules, which stands in contrast to the linear nature of DNA, RNA, and proteins. Combined with the large heterogeneity of their basic building blocks, the monosaccharides, they exhibit a significantly higher structural diversity than other abundant macromolecules. On the cell surface, carbohydrates (glycans) occur frequently as glycoconjugates, where they are covalently attached to proteins and lipids (aglycons). Glycosylation constitutes the most prevalent of all known post-translational protein modifications. It has been estimated that more than half the proteins in nature are glycoproteins [1]. Carbohydrates (*N*- or *O*-glycans) are typically connected to proteins via asparagine (*N*-linked glycosylation), serine or threonine (*O*-linked glycosylation). In recent years, it has been shown that glycosylation plays a key role in health and disease and consequently it has gained significant attention in life science research and industry [2–10].

Databases are playing a significant role in modern life science. It is now unthinkable to design research projects without a prior query or consultation of a few databases. In this respect, bioinformatics provides databases and tools to support glycobiologists in their research. Additionally, high throughput analysis of glycomes can only be handled properly with some sort of automated analysis pipeline that requires extensive bioinformatic support to organize the experimental data generated. In parallel, there are

---

M. Frank (✉) · S. Schloissnig  
Molecular Structure Analysis Core Facility-W160, Deutsches  
Krebsforschungszentrum (German Cancer Research Centre),  
Im Neuenheimer Feld 280, 69120 Heidelberg, Germany  
e-mail: m.frank@dkfz.de

bioinformatic groups actively developing mathematical or statistical algorithms and computational methodologies to analyze the data and thus uncover biological knowledge underlying the biological data. Since valuable experimental data are generated at various locations and in projects that target different scientific questions, the different sources of data have to be connected to generate a more complete data repository, which may aid in gaining a clearer understanding of the functions of carbohydrates in an organism. Consequently, data integration is a prerequisite for improving the efficiency of extraction and analysis of biological information, particularly for knowledge discovery and research planning [11].

Significant improvements in experimental analytical methods over recent years—particularly in glycan analysis by mass spectrometry and high performance separation techniques [12–21]—have led to a tremendous increase in the amount of carbohydrate structure data generated. The second source generating new experimental data on a large scale is the increased application of lectin and carbohydrate microarrays to probe the binding preferences of carbohydrates to proteins [22–26]. Consequently, the availability of databases and tools to store, retrieve, and analyze these data in an efficient way is of fundamental importance to progress in glycobiology [13, 15, 18, 27, 28]. Although bioinformatics for glycobiology or glycomics (‘glycoinformatics’) [29] is not yet as well established as in the fields of genomics and proteomics [30, 31], over the past few years, there has been a substantial increase in both the

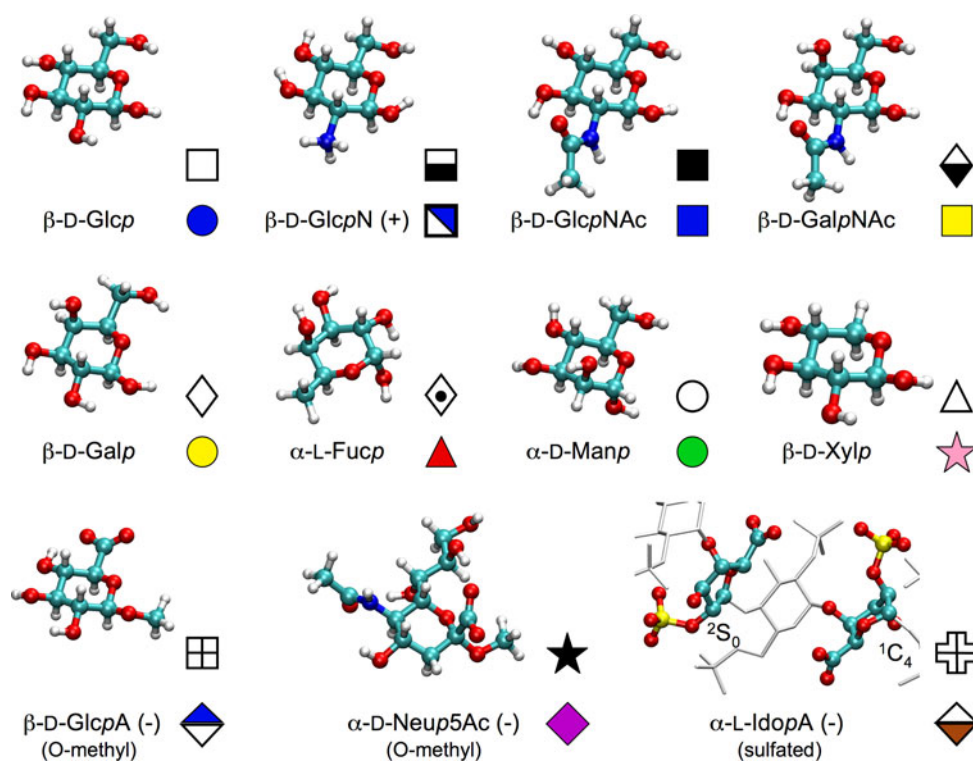
development, and use, of informatics tools and databases in glycosciences [32–42].

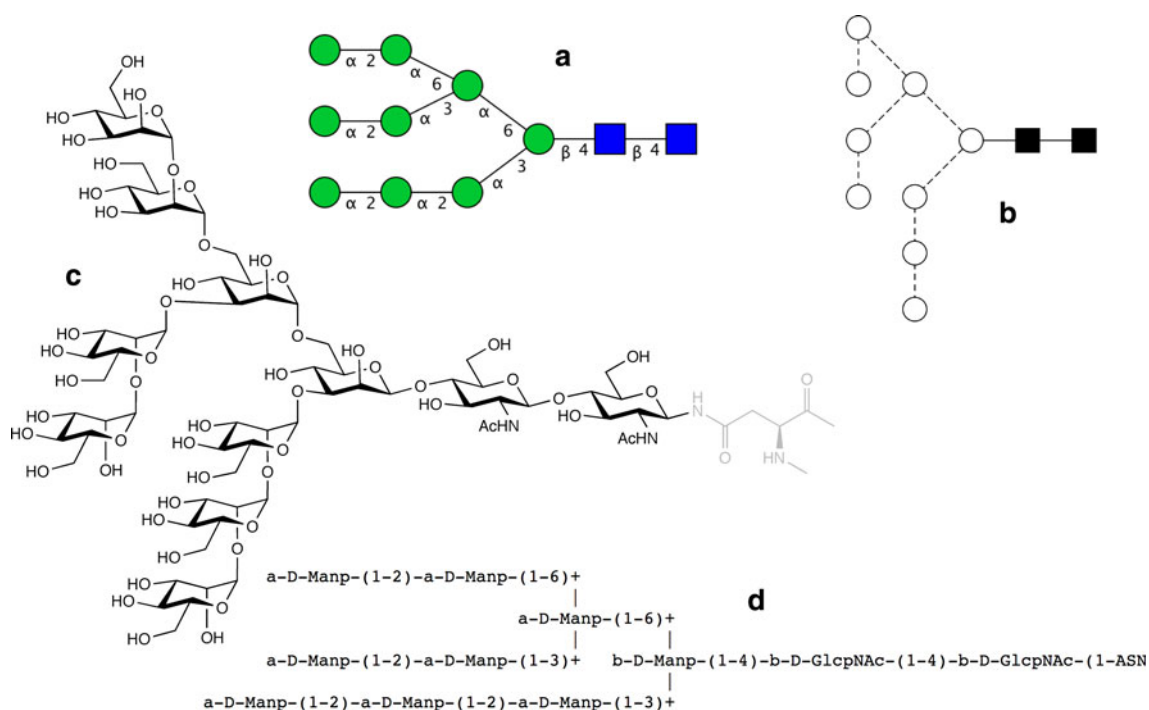
In this review, we will first introduce the various representations of carbohydrates used in the literature, then provide an overview of newly developed databases for glycomics, highlighting briefly the most recent glycoinformatic developments in sequence alignment and data mining, and provide an update [38] on tools to support experimental glycan analysis. Finally, we will review the field of (3D) structural glycoinformatics and molecular modeling of carbohydrates, glycoproteins, and protein–carbohydrate interaction.

## Graphical representations of carbohydrate structures

The basic units of carbohydrates are the monosaccharides. Whereas the other fundamental building blocks of biological macromolecules (nucleotides and amino acids) are clearly defined and limited in their number, the situation is much more complex for the carbohydrates. This becomes immediately evident by looking at the ten most frequently occurring monosaccharides in mammals [43]: D-GlcNAc, D-Gal, D-Man, D-Neu5Ac, L-Fuc, D-GalNAc, D-Glc, D-GlcA, D-Xyl, and L-IdoA (Fig. 1). Less than half of them are unmodified hexoses (D-Glc, D-Gal, D-Man) or pentoses (D-Xyl). Most of them are modified or substituted on the parent monosaccharides (deoxy: L-Fuc; acidic: D-GlcA, L-IdoA; substituted: D-GlcNAc, D-GalNAc, D-Neu5Ac).

**Fig. 1** Frequently occurring carbohydrate building blocks in mammalia. For each monosaccharide, the 3D structure, the IUPAC short code, and the symbols used in the Oxford (*top*) and CFG (*bottom*) symbolic nomenclature are shown. The acids (*last row*) are displayed with substituents (*O*-methyl, sulfate).  $\alpha$ -L-IdoA is shown in two conformations ( $^1C_4$  and  $^2S_0$ ) as they appear in heparin (pdb code 1E00 [53]). The formal charge of the monosaccharide at physiological pH is denoted in *parentheses*





**Fig. 2** Different graphical representations of the *N*-glycan  $\text{GlcNAc}_2\text{Man}_9$ . **a** CFG symbolic representation [45]. **b** Oxford system [44]. **c** Chemical drawing. **d** Extended IUPAC 2D graph representation

Therefore, derivatization is for monosaccharides the rule rather than an exception. The *D*-form is more common, but some monosaccharides occur more frequently in their *L*-form. Additionally, each of them can occur in two anomeric forms ( $\alpha/\beta$ ) and two ring forms [pyranose (*p*)/furanose (*f*)], which results, for example, in eight forms of cyclic galactose ( $\alpha$ -*D*-Galp,  $\alpha$ -*L*-Galp,  $\beta$ -*D*-Galp,  $\beta$ -*D*-Galf, etc.).

Simplified representations of complex biological macromolecules are frequently used to communicate or encode information on their structure. One-letter codes are in use to encode nucleic acids (5 nucleotides) or proteins (20 amino acids). Since the number of basic carbohydrate units frequently found in mammals is also very limited, symbolic representations [44, 45] (Figs. 1, 2) are frequently in use and one-letter codes have also been proposed [46, 47]. However, more than 100 different monosaccharides are found in bacteria, as has been revealed by a statistical database analysis [48]. This renders the general representation of monosaccharides by one-letter codes unfeasible, and generally longer abbreviations for the monosaccharide residues have to be used. A standardized International Union of Pure and Applied Chemistry (IUPAC) nomenclature for monosaccharides and oligosaccharide chains exists (<http://www.chem.qmul.ac.uk/iupac/2carb/>) [49], and full names and short codes for the common monosaccharides and derivatives have been defined (e.g., ‘Glc’ for glucose, ‘GlcNAc’ for *N*-acetylglucosamine). Typically

the short names are derived from the trivial names of the monosaccharides (e.g. ‘Fuc’, systematic name: 6-deoxygalactopyranose; trivial name: fucose). As already shown, in order to define the full monosaccharide short names, the anomeric descriptor, the *D/L* identifier and the ring form (*p/f*) have to be given as well, so the shortest name for ‘ $\alpha$ -*D*-glucopyranose’ would be ‘ $\alpha$ -*D*-Glc $p$ ’. An example of a more complex monosaccharide is *N*-acetyl- $\alpha$ -neuraminic acid; short name:  $\alpha$ -*D*-Neup5Ac or  $\alpha$ -Neu5Ac (full IUPAC name: 5-acetamido-3,5-dideoxy-*D*-glycero- $\alpha$ -*D*-galactono-2-ulopyranosonic acid).

The most commonly used graphical and textual representations for carbohydrates are shown in Fig. 2. Each of these shows a different level of information content that is tailored to a particular area of glycoscience research. Glycobiologists will prefer cartoon representations (Fig. 2a, b), whereas synthetic chemists will prefer the ‘chemical’ structural drawings with full atom topology displayed (Fig. 2c). Unfortunately, there are different graphical symbols in use for the same monosaccharides, which is even confusing for scientists working in the field. An agreement on one set of symbols would be very beneficial for the community [50]. From the viewpoint of bioinformatics, the graphical representations are only relevant for structure display in the context of user interfaces. Software tools have been developed that are able to generate on-the-fly cartoon representations from a carbohydrate sequence format (which is a ‘computer representation’ of a

carbohydrate structure) [51]. Although carbohydrates have been encoded successfully using the ‘computerized’ extended IUPAC (2D) representation [52] (Fig. 2d), it has been realized that a more flexible and systematic sequence format is required to encode all carbohydrate structures that occur in scientific publications.

## Bioinformatic concepts and algorithms

### Encoding of carbohydrate structures

There are essentially two possible ways to encode a carbohydrate molecule: as a set of atoms that are connected through chemical bonds, or as a set of building blocks that are connected through linkages. The first approach is used in chemoinformatics and a variety of chemical file formats (e.g., CML [54], InChi [55], SMILES [56]) have been developed for encoding of molecules for storage in chemical databases like PubChem [57] or ChEBI [58]. Figure 3 shows one PubChem entry of sialyl Lewis-X together with additional structural descriptors, like molecular weight. IUPAC full names and InChi and SMILES encoding are computed from the chemical drawing; therefore, encoding of carbohydrates as InChi or auto-generated IUPAC names is possible; however, there are severe limitations. One of the main requirements for databases is to store information in an organized way that facilitates further computational processing. Based on the

InChi or IUPAC code, it is, for example, very difficult to derive the monosaccharide composition of sialyl Lewis-X, which is  $(\text{Neu5Ac})_1(\text{Gal})_1(\text{GlcNAc})_1(\text{Fuc})_1$ . Additionally, since there is more than one entry for sialyl Lewis-X in the PubChem database, the automatic procedures to always generate the same InChi code for the same carbohydrate may need to be improved in order to generate unique IDs for carbohydrates. Although it is possible to develop software that parses InChi codes and assigns knowledge about monosaccharides to a database entry, InChi may not be the first choice for the encoding of carbohydrate structures. Consequently, it would be much more efficient to encode carbohydrates using a residue-based approach similar to the sequences of genes and proteins. However, there are two significant differences: the number of building blocks (residues) may be very large due to frequently occurring modifications of the parent monosaccharides, and the carbohydrate chains frequently contain branches, which means that many carbohydrates are tree-like molecules.

The prerequisite for a residue-based encoding format is a controlled vocabulary of the residue names. For practical reasons, the number of residues should be kept as small as possible. The main difficulty in encoding monosaccharide names in a systematic way is the definition of clear rules about which atoms of a molecule belong to a monosaccharide residue and which are, for example, of type ‘non-monosaccharide’ (e.g., substituent). The following short list will illustrate the dilemma: Glc, Gal, GlcN, GlcNAc, GalNAc, GlcOAc. For a biologist, all of them would be

**Fig. 3** PubChem entry of sialyl Lewis-X together with structural descriptors like IUPAC name and InChi and SMILES code

**Sialyl lewis-x - Compound Summary** (CID 9875818)

**Table of Contents**

- Synonyms
- Properties
- Descriptors
- Compound Information
- Substance Information
- Category
- Exports

**Depositor-Supplied Synonyms:** (Total: 12)

Sialyl lewis-x  
Slex  
Sialyl LeX  
C acid  
Sialyl Lewis X Mimetic  
Sialyl Lewis x (sLex)  
ydro-pyran-2-carboxylic acid  
CHEBI:111860  
CID9875818  
CID 4564901

**Properties Computed from Structure:**

Molecular Weight	820.74418 [g/mol]
Molecular Formula	C <sub>31</sub> H <sub>52</sub> N <sub>2</sub> O <sub>23</sub>
XLogP3-AA	-8.3
H-Bond Donor	15
H-Bond Acceptor	23

**Descriptors Computed from Structure:**

**IUPAC Name:** 5-acetamido-2-[2-[5-acetamido-6-hydroxy-2-(hydroxymethyl)-4-(3,4,5-trihydroxy-6-methylloxan-2-yl)oxyoxan-3-yl]oxy-3,5-dihydroxy-6-(hydroxymethyl)oxan-4-yl]oxy-4-hydroxy-6-[(1R,2R)-1,2,3-trihydroxypropyl]oxane-2-carboxylic acid

**Canonical SMILES:**  
CC1C(C(C(O1)OC2C(C(OC2OC3C(C(C(C(O3)CO)O)OC4(CC(C(C(O4)C(C(CO)O)NC(=O)C)O)C(=O)O)CO)O)NC(=O)C)O)O

**Isomeric SMILES:**  
CC1C(C(C(O1)OC2C(C(OC2OC3C(C(C(C(O3)CO)O)OC4(CC(C(C(O4)C(C@@H)C(C@@H)(CO)O)NC(=O)C)O)C(=O)O)CO)O)NC(=O)C)O)O

**InChI:** InChI=1S/C31H52N2O23/c1-8-17(41)20(44)21(45)28(50-8)54-25-16(33-10)(3)38)27(47)51-14(7-36)23(25)53-29-22(46)26(19(43)13(6-35)52-29)56-31(30(48)49)4-11(39)15(32-9(2)37)24(55-31)18(42)12(40)5-34/h8,11-29,34-36,39-47H,4-7H2,1-3H3,(H,32,37)(H,33,38)(H,48,49)/t87,117,12-,137,147,157,167,177,18-,197,207,217,227,237,247,257,267,277,287,297,317/m1/s1

**InChIKey:** NIGUVXFURDGGQKZ-TUNRMZMNSA-N

monosaccharides on their own (glucose, galactose, glucosamine, *N*-acetylglucosamine, etc.) except GlcOAc, which would be a glucose carrying an ‘acetyl’ substituent. From an encoding point of view, it makes more sense that all entries are of type ‘Glc’ or ‘Gal’ and N, NAc, and OAc are substituents, respectively. Even for bacterial monosaccharides, this would result in a reasonably sized ‘monosaccharide base type’ residue list {Glc, Gal, ...} and ‘non-monosaccharide’ residue list {N, NAc, OAc, ...}, which is much easier to maintain.

Over the years, each database project that aimed to store carbohydrate structures developed a new sequence format (see Table 1). Because of this, translation tools for the different formats were necessary to establish cross-links between the databases. In the context of the EUROCarbDB design study, which aimed at developing standards for glycoinformatics, a new carbohydrate sequence format for use in databases (GlycoCT) [59] has been established as well as a reference database for monosaccharide notation (<http://www.monosaccharidedb.org>). The GlycomeDB [60] project uses GlycoCT as a standard format for the integration of all open access carbohydrate structure databases (Fig. 4). Recently, as a result of a close collaboration between developers at the Complex Carbohydrate Research Center (CCRC) and EUROCarbDB, the Glyde-II format (<http://glycomics.ccr.c.uga.edu/GLYDE-II/>) was created where concepts of Glyde [61] and GlycoCT [59] were combined in order to define a new standard exchange format for carbohydrate structures [27].

#### Algorithms for structural alignment and similarity of carbohydrate structures

For many applications in glycoinformatics, it is required to classify glycans based on occurring structural motifs or patterns, or to compare two carbohydrates and to establish

a similarity score. Despite many efforts, this is still an open problem due to the lack of broadly accepted metrics on carbohydrate structures. Algorithms based on adapting the already established sequence alignment approaches from DNA, RNA, and protein sequences to carbohydrates and establishing a scoring matrix for substitutions have been proposed [62, 63]. However, discovery of biomarkers or more broadly extracting discriminating patterns from sets of carbohydrates poses a great challenge due to their branched nature and the possibility that a significant pattern can be fragmented and distributed across multiple branches of the carbohydrate. Traditional pattern discovery and classification techniques from machine learning have been applied with increasing success to meet this challenge. Markov models were used to discover patterns spanning multiple branches [62, 63]. Then the focus shifted to Support Vector Machines and the search for kernels appropriate for branched structures [64–66]. Recently, a novel statistical approach for motif discovery that currently outperforms all competing methods has been presented [67].

#### Glycomics ontologies

The information explosion in biology makes it difficult for researchers to stay up-to-date with current biomedical knowledge and to make sense of the massive amounts of online information. Ontologies are increasingly enabling biomedical researchers to accomplish these tasks [68]. An ontology provides a shared vocabulary, which can be used to model a domain of interest, and it defines the type of objects and concepts that exist, and their properties and relations. Ontologies are often represented graphically as a hierarchical structure of concepts (nodes) that are connected by their relationships (edges). Concepts and relationships are assigned unique ontological names. For

**Table 1** Major carbohydrate structure databases and the sequence formats used

Database	Encoding	URL
GlycomeDB [89]	GlycoCT [59]	<a href="http://www.glycome-db.org/">http://www.glycome-db.org/</a>
EUROCarbDB <sup>a</sup>	GlycoCT [59]	<a href="http://www.ebi.ac.uk/eurocarb/">http://www.ebi.ac.uk/eurocarb/</a>
CarbBank <sup>a</sup> [52]	IUPAC extended [90]	<a href="http://www.boc.chem.uu.nl/sugabase/carbbank.html">http://www.boc.chem.uu.nl/sugabase/carbbank.html</a>
KEGG <sup>a</sup> [83]	KCF [91]	<a href="http://www.genome.jp/kegg/glycan/">http://www.genome.jp/kegg/glycan/</a>
GLYCOSCIENCES.de <sup>a</sup> [82]	LINUXS [92]	<a href="http://www.glycosciences.de/">http://www.glycosciences.de/</a>
CFG <sup>a</sup> [84]	Glycominds Linear Code <sup>®</sup> [47]	<a href="http://www.functionalglycomics.org/">http://www.functionalglycomics.org/</a>
BCSDB <sup>a</sup> [93]	BCSDB linear code	<a href="http://www.glyco.ac.ru/bcsdb3/">http://www.glyco.ac.ru/bcsdb3/</a>
GlycoSuiteDB [87]	IUPAC condensed [94]	<a href="http://glycosuitedb.expasy.org/">http://glycosuitedb.expasy.org/</a>
GlycoBase (Dublin) <sup>a</sup> [86]	Motif based	<a href="http://glycobase.ucd.ie/">http://glycobase.ucd.ie/</a>
GlycoBase (Lille) <sup>a</sup> [95]	Linkage path	<a href="http://glycobase.univ-lille1.fr/base/">http://glycobase.univ-lille1.fr/base/</a>
JCGGDB [96]	CabosML [97]	<a href="http://jcgddb.jp/">http://jcgddb.jp/</a>

<sup>a</sup> Currently queried by GlycomeDB

**Fig. 4** GlycomeDB entry of sialyl Lewis-X. The structure is displayed in a ‘human readable’ encoding (CFG cartoons) at the *top* and in a ‘computer readable’ sequence format (GlycoCT) at the *bottom*. Additionally, links to external databases and information on structural motifs are available

The screenshot displays the GlycomeDB interface for structure 7.552. On the left is a navigation menu with options like Home, Database, Search by database ID, Exact structure search, Substructure search, Similarity search, MCS search, Search by species, Translation, Download, Documentation, and Contact. The main content area is titled 'Structure information for structure 7.552' and includes a 'Picture' section with a CFG cartoon (a diamond, a circle, and a square connected by lines with labels like α 3, β 4, and α 3) and a 'Species' section listing '9606 Homo sapiens'. Below that is a 'Remote structures' table with columns for database name and count. The 'Motifs' section lists Lewis X, Sialyl Lewis X, Lactosamine motif, and VIM. The 'Structure encodings' section has a 'GlycoCT(XML)' button and a 'Get Encoding' button. The 'GlycoCT' section shows the sequence: RES, 1b:x-dg1c-HEX-1:5, 2s:n-acetyl, 3b:a-1gal-HEX-1:5|6:d, 4b:b-dgal-HEX-1:5, 5b:a-dgro-dgal-NON-2:6|1:a|2:keto|3:d, 6s:n-acetyl, LIN, 1:1d(2+1)2n, 2:1o(3+1)3d, 3:1o(4+1)4d, 4:4o(3+2)5d, 5:5d(5+1)6n.

Database	Count
Carbank (CCSD)	46377
Carbank (CCSD)	46708
CFG	carbOlink_47412_D000
EUROCarbDB (EBI)	6939
EUROCarbDB (NIBRT)	6939
GLYCOSCIENCES.de	11891
KEGG	G00242

example, the ontological concepts ‘carbohydrate’ and ‘molecule’ can be connected by the hierarchical relationship ‘is\_a’. In this way, ontologies can be used to provide a formal mechanism to categorize objects by specifying their membership in a specific class. The Glycomics Ontology (Glyco) focuses on the glycoproteomics domain to model the structure and functions of glycans and glycoconjugates, the enzymes involved in their biosynthesis and modification, and the metabolic pathways in which they participate. Glyco is intended to provide both a schema and a sufficiently large knowledge base, which will allow classification of concepts commonly encountered in the field of glycobiology in order to facilitate automated information analysis in this domain [69, 70] (for more information, see the web site <http://lstdis.cs.uga.edu/projects/glycomics/>).

#### Predicting the size and diversity of glycomes

A hexasaccharide can, in theory, build  $10^{12}$  structural isomers [71]. Although such large numbers highlight the intrinsic complexity of carbohydrates, the actual number of carbohydrates in nature is probably significantly smaller. Modeling of enzyme kinetics and estimating the size and

diversity of the glycome has been of considerable interest to the field of glycobiology. The first attempt to establish mathematical models of N-type glycosylation occurred over 10 years ago [72], by employing the known enzymatic activities of glycosyltransferases involved in the N-type glycosylation pathway and generating all possible carbohydrates resulting from those activities up to the first galactosylation of an oligosaccharide. This work was then extended [73] by using less simplifying assumptions and extending the set of enzymes included in the model. Recent advances have led to the ability to estimate enzyme reaction rates and enzyme concentrations from mass spectrometry data, thereby opening up the possibility to infer changes to the enzyme concentrations in diseased tissues [74]. Similarly, the expression profiles of glycosyltransferases were used to predict the repertoire of potential glycan structures [75, 76]. A recent estimation of the size of the human glycome based on biosynthetic pathway knowledge approximates the upper limit of distinct carbohydrates to be in the range of hundreds of thousands [77]. Currently, there are about 35,000 distinct carbohydrate structures stored in databases [60]; however, nobody knows how many carbohydrate structures have already been discovered or published so far, since no

comprehensive database of carbohydrate structures exists. Statistical analyses of the structures stored in GLYCOSCIENCES.de and the Bacterial Carbohydrate Structure Database (BCSDB) have been performed and showed the differences between available structures from the mammalian and bacterial glycome [48]. The ‘chemical analysis’ of the database entries revealed that about 36 chemical building blocks (=monosaccharide + branching level) would be required for the chemical synthesis of 75% of the mammalian glycans [43]. Unfortunately, the situation is much more complex for the bacterial carbohydrates. However, if one would concentrate on the synthesis of a particular subclass of bacterial carbohydrates (e.g., for vaccine development) the diversity becomes much smaller.

### Databases and tools for glycobiology

A variety of databases are available to the glycoscientist [32, 36, 78–80]. From the viewpoint of glycoproteins, they can be grouped into databases that contain information on the proteins themselves, databases that store information on the enzymes and pathways that build the glycans, and carbohydrate structure databases [36, 37]. Only limited information is available in databases on glycoforms of glycoproteins.

#### Carbohydrate structure databases

The complex carbohydrate structure database (CCSD)—often referred to as *CarbBank* in reference to its query software—was developed and maintained for more than 10 years by the Complex Carbohydrate Research Center of the University of Georgia (USA) [52, 81]. The CCSD was the largest effort during the 1990s to collect the structures of carbohydrates, mainly through retrospective manual extraction from literature. The main aim of the CCSD was to catalog all publications in which complex carbohydrate structures were reported. Unfortunately, funding for the CCSD stopped during the second half of the 1990s and the database was no longer updated. Nevertheless, with almost 50,000 records (from 14,000 publications) relating to approximately 23,500 different carbohydrate sequences, the CCSD is still one of the largest repositories of carbohydrate-related data. Subsets of the CCSD data have been incorporated into almost all recent open access databases, of which the major ones are GLYCOSCIENCES.de (23,233) [82], KEGG Glycan database (10,969) [83], CFG Glycan Database (8,626) [84], Bacterial Carbohydrate Structure Database (BCSDB, 6,789) [85], and GlycoBase (377) [86]. The numbers in parentheses denote the number of distinct carbohydrate sequences (without aglycons) stored in the database (based on GlycomeDB analysis,

October 2008). Recently, the JCGGDB, which assembles CabosDB, Galaxy, LipidBank, GlycoEpitope, LfDB, and SGCAL (with 1,490 unique carbohydrate structures) and the GlycoSuiteDB (with about 3,300 unique carbohydrate structures) [87] have become freely accessible as well.

The EUROCarbDB project (<http://www.eurocarb.org>) was a design study that aimed to create the foundations for a new infrastructure of distributed databases and bioinformatics tools where scientists themselves can upload carbohydrate structure-related data. Fundamental ethics of the project were that all data are freely accessible and all provided tools are open source. A prototype of a database application has been developed that can store carbohydrate structures plus additional data such as biological context (organism, tissue, disease, etc.), and literature references. Primary experimental data (MS, HPLC, and NMR) that can serve as evidence or reference data for the carbohydrate structure in question can be uploaded as well (Fig. 5).

Until recently, there was hardly any direct cross-linking between the established carbohydrate databases [88]. This is mainly due to the fact that the various databases use different sequence formats to encode carbohydrate structures [59] (Table 1). Therefore, the situation in glycoinformatics has been characterized by the existence of multiple disconnected and incompatible islands of experimental data, data resources, and specific applications, managed by various consortia, institutions, or local groups [27, 37]. Importantly, no comprehensive and curated database of carbohydrate structures currently exists. From the user’s point of view, the lack of cross-links between carbohydrate databases means that, until recently, they had to visit different database web portals in order to retrieve all the available information on a specific carbohydrate structure. Additionally, the users might have had to acquaint themselves with the different local query options, some of which require knowledge of the encoding of the residues in the respective database.

In 2005, a new initiative was begun to overcome the isolation of the carbohydrate structure databases and to create a comprehensive index of all available structures with references back to the original databases. To achieve this goal, most structures of the freely available databases were translated to the GlycoCT sequence format [59], and stored in a new database, the GlycomeDB [60]. The integration process is performed incrementally on a weekly basis, updating the GlycomeDB with the newest structures available in the associated databases. During the integration process, some automated checks are performed. Structures that contain errors are reported to the administrators of the original database. A web interface has been developed (<http://www.glycome-db.org>) as a single query point for all open access carbohydrate structure databases [89] (Fig. 4).

The figure shows two panels of the EuroCarbDB web interface. Panel (a) is the GlycanBuilder Tool, which includes a menu bar (File, Edit, Structure, View, Help), a toolbar with various icons, and a glycan structure diagram. Below the diagram, there is an input field for 'm/z: 1345,6723 [MONO,perMe,Na,0,freeEnd]'. Panel (b) is the search results page, titled 'Search structures'. It shows a search query 'substructure is given at Anywhere' and 'Found 808' results. A table lists the results with columns: Structure, Entered, Contributor, Data, and Taxonomies. The first three entries have associated MS data (MS) and are from 'Bothrops moojeni'. The fourth entry is from 'Sus scrofa'.

Structure	Entered	Contributor	Data	Taxonomies
	22.04.2008	maass	MS	Bothrops moojeni
	22.04.2008	maass	MS	Bothrops moojeni
	22.04.2008	maass	MS	Bothrops moojeni
	21.07.2008	guest	-	Sus scrofa

**Fig. 5** EUROCarbDB web interface. **a** The GlycanBuilder Tool [51] serves as an interface for structure input. Various graphical representations are supported and can be changed interactively.

**b** Result of a structure search in the database. The first three entries contain associated MS data as evidence for the structure

#### Databases for carbohydrate–protein interaction data

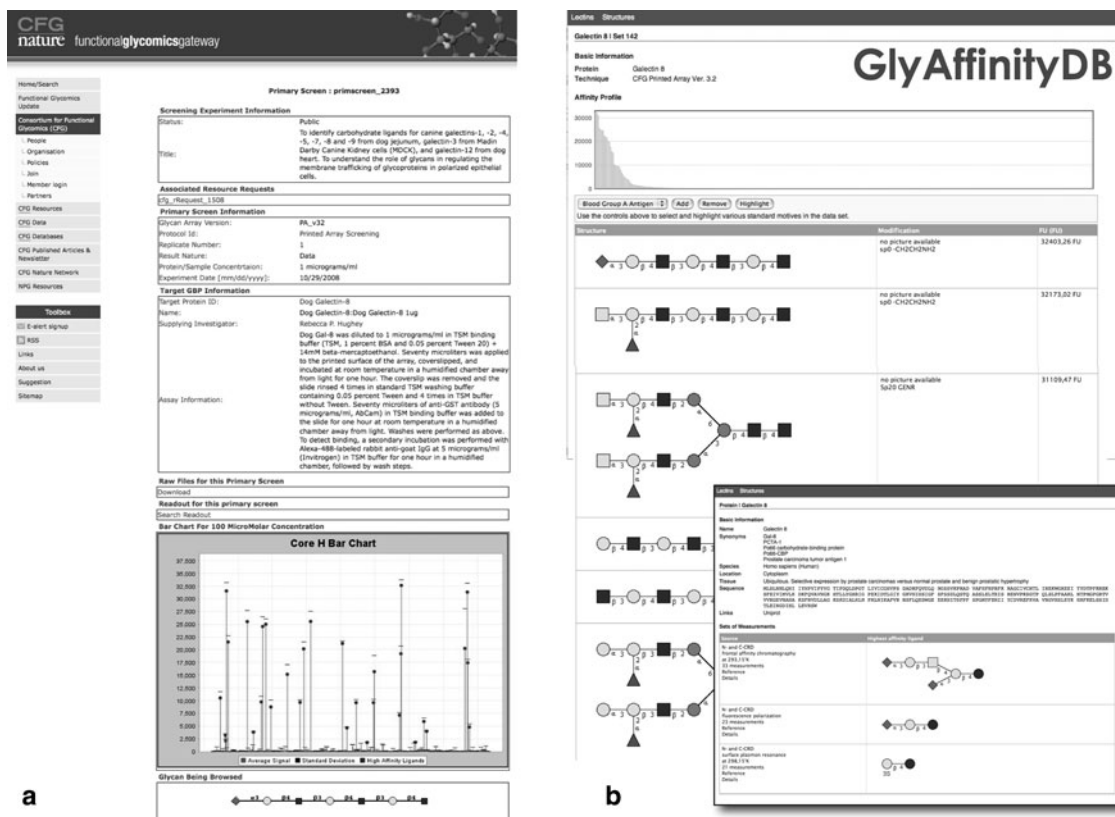
Advances in recent years have led to an explosive growth of data from carbohydrate microarray experiments, coming from multiple research laboratories, each employing their own proprietary technology for spotting the array [22, 98–101]. Unfortunately, the often radically different approaches to the spotting of the arrays can change the binding affinities observed. This inhomogeneity in the way the data are generated has caused problems in the comparative analysis and evaluation of the data. Further complications arise from lack of comprehensive applications to manage the data generated in the experiments and also computational approaches to perform analytical studies. There is a clear need in this area for more research and development of computational approaches and tools. Particularly, the standards for reporting glycan array experiments need to be defined [102].

Currently, there is one major public resource for glycan array data provided by the Consortium for Functional Glycomics (CFG) (<http://www.functionglycomics.org>). The CFG supported the development of the first robot-produced, publicly available micro-titer-based glycan array. The currently used printed mammalian glycan microarray format (version 4.1) comprises 465 synthetic and natural glycan sequences representing major glycan structures of glycoproteins and glycolipids [103]. In 2008, a pathogen glycan array was also made available for

screening, containing 96 polysaccharides derived from Gram-negative bacteria. The protein–glycan interaction core (H) analyzes investigator-generated lectins, antibodies, antisera, microorganisms, or suspected glycan binding proteins (GBP) of human, animal, and microbial origins on the mammalian and pathogen glycan microarrays. Fluorescent reagents are used for detecting primary binding to the glycans on the array. The results of the screening performed by the Core H can be accessed through the consortium web page, and the raw data can be downloaded as an Excel spreadsheet [84] (Fig. 6). The website offers an interactive bar-chart that dynamically displays the glycan structures upon a mouse click on the signal of interest. Links to CFG databases that contain curated information on the GBP and the glycan structures printed on the array are available.

A second large resource of carbohydrate–protein interaction data is the Lectin Frontier DataBase (LfDB) provided by the Japanese Consortium for Glycobiology and Glycotechnology (JCGG). A significant part of the data may have been generated as part of the structural glycomics project funded by New Energy and Industrial Technology Organization (NEDO) [104]. In contrast to the CFG glycan microarray database, which provides relative fluorescence units (FU), the LfDB provides affinity constants ( $K_a$ ) determined by frontal affinity chromatography [24, 105]. Similar to the CFG website, users can navigate to the experimental data using the lectin as an entry point. The





**Fig. 6** Glycan microarray data provided by the Consortium for Functional Glycomics (CFG). **a** Web interface providing access to the primary data and related information. **b** The CFG dataset in the

GlyAffinity database. Affinity data from different techniques (FAC, SPR, FP) are also available

data are also presented as an interactive bar-chart (Fig. 7). Since the site is to a large extent in Japanese, it is somewhat difficult to explore the full functionality of the web-interface at the moment.

Recently, a prototype for an integrated database for protein–carbohydrate interaction, ‘GlyAffinity’, has been developed at the German Cancer Research Center. The database aims at providing a comprehensive repository of curated protein–carbohydrate interaction data from various sources and techniques (Figs. 6b and 7b). The publicly available microarray experiments conducted by the CFG have been acquired and processed. This entailed the parsing of the data, conversion of the carbohydrate structures to the GlycoCT format, and curation of an initial set of approximately 100 experiments for inclusion in the database. The complete contents of the Lectin Frontier DataBase have also been imported and curated, and publications have been scoured for data and manually entered. The Leffler Laboratory (Lund, Sweden) provided access to their primary data, which has been processed and partially imported, and initial steps have been taken to include the data generated by the Feizi Laboratory (London, UK). Access to the data is provided through an interactive web-

interface, which offers options to locate data either by lectin or carbohydrate structure. Each lectin entry is classified according to the established hierarchical lectin family scheme [106–108] and provides a list of experiments conducted with their experimental conditions and technique. The pages of the individual experiments feature the full list of carbohydrates, their recorded affinity, and the possibility to detect standard motifs.

A current limitation in making full use of protein–carbohydrate interaction data is the lack of systematic analysis methods for extracting information, most importantly the deduction of the binding epitope. Recently, the development of a novel algorithm to detect the occurrence of significant motifs in carbohydrate microarray experiments has been reported [109]. The approach entails the selection of 63 commonly occurring carbohydrate motifs (e.g., Lewis-X, terminal beta-GalNAc, etc.) and processing the complex carbohydrates found on the CFG microarray to detect their presence. Subsequent analysis of the occurrences of a motif in the carbohydrate structures of a particular experiment together with the fluorescence intensity measured yields information about the specificity the lectin exhibits towards a subset of the motifs.



**Table 2** Protein databases that contain carbohydrate related information (see also [32])

Name	Content	URL
CAZy [110]	Carbohydrate active enzymes	<a href="http://www.cazy.org/">http://www.cazy.org/</a>
BRENDA [113]	Enzymes	<a href="http://www.brenda-enzymes.org/">http://www.brenda-enzymes.org/</a>
JCGGDB [96]	Glyco genes, glycoproteins, lectin affinity, MS data, epitopes	<a href="http://jcgddb.jp/">http://jcgddb.jp/</a>
CFG [84]	Glycans, lectins, glycosyltransferases	<a href="http://www.functionalglycomics.org/">http://www.functionalglycomics.org/</a>
Glyco3D [115, 120]	3D structures of carbohydrates, glycosyltransferases and lectins	<a href="http://www.cermav.cnrs.fr/glyco3d/">http://www.cermav.cnrs.fr/glyco3d/</a>
O-GlycBase [117]	Curated set of <i>O</i> -and <i>C</i> -glycosylated proteins	<a href="http://www.cbs.dtu.dk/databases/OGLYCBASE/">http://www.cbs.dtu.dk/databases/OGLYCBASE/</a>
UniProt/SWISS-PROT [121]	Annotated proteins (lectins, glycoproteins, enzymes)	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>
RCSB Protein Data Bank [112]	3D structures of lectins, glycoproteins, enzymes	<a href="http://www.rcsb.org/pdb/">http://www.rcsb.org/pdb/</a>

information resource on glycoproteins is GlycoSuiteDB [87] cross-linked with UniProt/SWISS-PROT [119].

### Software tools for glycan analysis

Over the years, the increased application of a variety of methods for glycan analysis have led to the development of many software tools that aim to assist in the interpretation of the experimental data generated. High-throughput methods are mainly based on mass spectrometry (MS) and high performance liquid chromatography (HPLC) due to their sensitivity. Many of these methods rely on reference data of known carbohydrates from databases. Nuclear magnetic resonance (NMR) spectroscopy has always played a key role in the de novo determination of carbohydrate structures [122], and some new tools that aim at decreasing the time needed for NMR peak assignment have been recently developed (Table 3).

Automatic interpretation of MS data remains a challenging task due to the technique only being able to report masses of the fragments observed. This information can be

used to calculate a set of potential compositions the carbohydrate could have, which then give rise to a set of potential carbohydrates that satisfy the composition constraints. One of the first tools to deduce potential carbohydrate compositions from MS data is the GlycoMod online service [123]. The number of compositions matching a certain mass value scales exponentially with the number of different monomers; therefore, taxonomic and biosynthetic information have to be incorporated in the assignment process so that smart choices can be made based upon a given composition. The Cartoonist tool [124–127] uses a set of archetypal structures in combination with a set of rules for their potential modification to generate all types of glycans that could be possibly synthesized by mammalian cells. Archetypes and rules have been compiled by a group of experts and represent the current knowledge on biosynthetic pathways in mammalian organisms. The number of matching structures is thus greatly reduced by avoiding implausible molecules. Cartoonist has been extended over the years, and now uses a well-tested library of structures, scoring routines to establish confidence scores in an annotation, and knowledge of

**Table 3** New free software tools for glycan analysis (see also [38])

Name	Content	URL
GlycoWorkbench [136]	Glycan MS spectra annotation	<a href="http://www.eurocarbdb.org/applications/ms-tools">http://www.eurocarbdb.org/applications/ms-tools</a> <a href="http://www.ebi.ac.uk/eurocarb/gwb/home.action">http://www.ebi.ac.uk/eurocarb/gwb/home.action</a>
Glyco-Peakfinder [132]	Determination of glycan compositions from their mass signals	<a href="http://www.eurocarbdb.org/applications/ms-tools">http://www.eurocarbdb.org/applications/ms-tools</a> <a href="http://www.glyco-peakfinder.org/">http://www.glyco-peakfinder.org/</a>
GlycoPep ID [142]	Identifying the peptide moiety of glycopeptides generated using a nonspecific enzyme	<a href="http://hexose.chem.ku.edu/predictiontable.php">http://hexose.chem.ku.edu/predictiontable.php</a>
GlycoMiner [138]	Glycopeptide composition analysis	<a href="http://www.chemres.hu/ms/glycominer/index.php">http://www.chemres.hu/ms/glycominer/index.php</a>
AutoGU [86]	HPLC analysis	
ProspectND	NMR spectra processing	<a href="http://www.eurocarbdb.org/applications/nmr-tools">http://www.eurocarbdb.org/applications/nmr-tools</a>
CCPN Tools [148]	Annotation of NMR spectra	<a href="http://www.ccpn.ac.uk/">http://www.ccpn.ac.uk/</a>
CASPER [146]	Structure determination of oligosaccharides and regular polysaccharides	<a href="http://www.eurocarbdb.org/applications/nmr-tools">http://www.eurocarbdb.org/applications/nmr-tools</a> <a href="http://relax.organ.su.se:8123/eurocarb/casper.action">http://relax.organ.su.se:8123/eurocarb/casper.action</a>

biochemical pathways to aid the interpretation of spectra. Cartoonist is used as an analysis tool for the glycan profiling service of the CFG.

De novo sequencing tools based on MS<sup>n</sup> fragmentation data are STAT [128], OSCAR [129] StrOligo [130], and GLYCH [131]. Glyco-Peakfinder [132] is a new web-service for the de novo determination of the composition of glycan-derived MS signals independent of the source of spectral data. Library-based sequencing methods—similar to those applied in proteomics—are also applied in glycomics, where experimental peaks of MS<sup>2</sup> spectra are compared with online calculated theoretical fragments from user-definable carbohydrate sequences deposited in databases (glyco-fragment mass fingerprinting). Such an approach is supported by the GlycoSearchMS service [133], where Glyco-Fragment [134] is used to calculate fragment-libraries of all carbohydrates contained in GLY-COSCIENCES.de. GlycosidiQ<sup>TM</sup> [135] is a similar mass fingerprinting tool developed for interpretation of oligosaccharide mass spectrometric fragmentation based on matching experimental data with theoretically fragmented oligosaccharides generated from the database GlycoSuiteDB [87]. However, the success of such an approach depends on the comprehensiveness of experimentally determined glycan structures included in the database. GlycoWorkbench [136], one of the most recent tools for the computer-assisted annotation of mass spectra of glycans, has been developed as an open source project in the context of the EUROCarbDB project. The main task of GlycoWorkbench is to evaluate a set of structures proposed by the user by matching the corresponding theoretical list of fragment masses against the list of peaks derived from the spectrum. For annotation, GlycoWorkbench uses a database of carbohydrate structures derived from GLY-COSCIENCES.de, CarbBank, and CFG glycan. The tool provides an easy to use graphical interface, a comprehensive and increasing set of structural constituents, an exhaustive collection of fragmentation types, and a broad list of annotation options. Mass spectra annotated with GlycoWorkbench can be uploaded into the EUROCarbDB MS database. GlyQuest [137] and GlycoMiner [138] are new tools that support high-throughput composition and primary structure determination of *N*-glycans attached to peptides, based on CID (collision induced dissociation) MS/MS (tandem mass spectrometric data). Also new is SysBioWare [139], a general software platform for carbohydrate assignment based on MS data. More specialized applications have been reported for the analysis of glycosaminoglycans [140], in silico fragmentation of peptides linked to *N*-glycans [141], and identifying the peptide moiety of sulfated or sialylated carbohydrates [142]. Of further interest for bioinformatic developers might be the OpenMS [143] initiative that provides an open source

framework for mass spectrometry and TOPPView [144], an open source viewer for MS data.

The development of robust high-performance liquid chromatography (HPLC) technologies continues to improve the detailed analysis and sequencing of glycan structures released from glycoproteins. In the context of the EUROCarbDB project, an analytical tool (autoGU) [86] was developed to assist in the interpretation, assignment, and annotation of HPLC-glycan profiles. AutoGU assigns provisional structures to each integrated HPLC peak and, when used in combination with exoglycosidase digestions, progressively assigns each structure automatically based on the footprint data. The software is assisted by GlycoBase, a relational database originally developed at the Oxford Glycobiology Institute, which contains the HPLC elution positions for over 350 2-AB labelled *N*-glycan structures together with digestion pathways. The system is suitable for automated analysis of *N*-linked sugars released from glycoproteins and allows detection of the carbohydrates at femtomolar concentrations [18].

ProSpectND is an advanced integrated NMR data processing and inspection tool, originally developed at the University of Utrecht and refined during the EUROCarbDB project. It allows batch processing of spectra simulations and automated graphics generation. CASPER is a tool for calculating chemical shifts of oligo- and polysaccharides as well as *N*- and *O*-glycans. The tool already has a long history [145, 146] and has recently received a major upgrade. CASPER can also be used for determining the primary structures of carbohydrates by simulating spectra of a set of possible structures and comparing them with the supplied experimental data to find the best match. The Collaborative Computing Project for the NMR community (CCPN) has developed a powerful data model for NMR experiments [147] and an assignment package for NMR spectra of proteins and peptides (CcpNmr Analysis) [148]. As a result of an intensive EUROCarbDB/CCPN collaboration, support for (branched) carbohydrates was added to the CCPN data model and CcpNmr Analysis software. Additionally, CASPER can be used to automatically assign NMR signals to carbohydrate atoms in connection with CcpNmr Analysis. CCPN project files can be directly uploaded into the EUROCarbDB NMR database.

#### Prediction and statistical analysis of glycosylation sites

Glycosylation is the most common post-translational modification of proteins [1]. Initial analyses yielded a consensus sequence motif for *N*-type glycosylation, Asn-X-Ser/Thr, with any amino acid at X except proline. Every *N*-type glycosylation site adheres to this motif, but its sole presence in the amino acid sequence is only a necessity and not sufficient to predict the presence of a glycan. The

situation becomes further complicated when one considers mucin-type or other types of *O*-glycosylation where the glycan is usually attached to a serine or threonine. Research in this area over the last 10 years has resulted in a series of approaches for the prediction of glycosylation sites (Table 4). All strategies for the prediction of glycosylation sites are of a statistical nature: NetCGlyc [149], NetNGlyc [150], NetOGlyc [151], and YinOYang [150] all use neural networks for the prediction of glycosylation sites; big-Pi [152] employs scoring functions based on amino acid properties; GPI-SOM [153] uses a Kohonen map; CKSAAP\_OGlySite [154], and EnsembleGly [155] use a Support Vector Machine based approach; and GPP [156], the currently best performing prediction tool, uses a hybrid combinatorial and statistical learning approach based on random forests. Training datasets for the statistical learning approaches are usually derived from the PDB or O-GLYCBASE [117].

The mechanism of acceptor site selection for the covalent attachment of carbohydrates by a series of glycosyl transferases in the case of C-type and O-type glycosylation, and by the oligosaccharyltransferase (OST) for N-type glycosylation, is still not completely understood. This has led to studies on the statistical and structural properties of amino acids in the neighborhood of glycosylation sites. Conformational and statistical properties of N-type glycosylation sites were analyzed based on glycosylated proteins found in the PDB [157]. Statistical properties of O-type glycosylation were analyzed based on entries of the O-GLYCBASE [158]. The GlySeq and GlyVicinity online services [159] allow for interactive exploration of the statistical and conformational properties of glycosylation sites and their surroundings.

## Carbohydrate 3D structures and molecular modeling

Complex carbohydrates represent a particularly challenging class of molecules in terms of describing their three-dimensional (3D) structure. Due to their inherent flexibility, these molecules very often exist in solution as an ensemble of conformations rather than as a single well-defined structure. Traditionally, NMR methods, especially Nuclear Overhauser Effect (NOE) measurements, have been widely used to study oligosaccharide conformation in solution [160, 161]. Unfortunately, many oligosaccharide NOEs cannot be resolved or are difficult to assign. Additionally, there are often too few inter-residue NOEs to make an unambiguous 3D structure determination possible. In general, the interpretation of structural experimental data frequently needs to be supported by molecular modeling methods [162, 163]. One of the main aims of computer modeling of carbohydrates is to generate reasonable 3D models that can be used to rationalize experimentally derived observations. Conformational analysis by computational methods consequently plays a key role in the determination of 3D structures of complex carbohydrates.

In recent years, a variety of modeling methods have been applied to the conformational analysis of carbohydrates [164]. Of these, the calculation of conformational maps for disaccharides using systematic search methods, and molecular dynamics (MD) simulations of oligosaccharides in explicit solvent, are by far the most popular methods in modeling of carbohydrate 3D structures [165–168]. Although quantum mechanics (*ab initio*) methods are used for modeling of carbohydrate conformation, these methods are still computationally too demanding to be used routinely to study or predict the 3D structure of complex

**Table 4** Tools for prediction and analysis of glycosylation sites

Name	Description	URL
Big-Pi [152]	GPI-anchors	<a href="http://mendel.imp.ac.at/sat/gpi/gpi_server.html">http://mendel.imp.ac.at/sat/gpi/gpi_server.html</a>
GPI-SOM [153]	GPI-anchors	<a href="http://gpi.unibe.ch/">http://gpi.unibe.ch/</a>
NetCGlyc [149]	C-mannosylation	<a href="http://www.cbs.dtu.dk/services/NetCGlyc/">http://www.cbs.dtu.dk/services/NetCGlyc/</a>
NetNGlyc [150]	N-glycosylation	<a href="http://www.cbs.dtu.dk/services/NetNGlyc/">http://www.cbs.dtu.dk/services/NetNGlyc/</a>
NetOGlyc [151]	O-glycosylation	<a href="http://www.cbs.dtu.dk/services/NetOGlyc/">http://www.cbs.dtu.dk/services/NetOGlyc/</a>
YinOYang [150]	O-beta-GlcNAc-ylation	<a href="http://www.cbs.dtu.dk/services/YinOYang/">http://www.cbs.dtu.dk/services/YinOYang/</a>
EnsembleGly [155]	O-, N- and C-glycosylation	<a href="http://turing.cs.iastate.edu/EnsembleGly/">http://turing.cs.iastate.edu/EnsembleGly/</a>
CKSAAP_OGlySite [154]	Mucin-type O-glycosylation	<a href="http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlySite/">http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlySite/</a>
GPP [156]	O- and N-glycosylation	<a href="http://comp.chem.nottingham.ac.uk/glyco/">http://comp.chem.nottingham.ac.uk/glyco/</a>
GlySeq [159]	Statistical analysis of glycosylation sites based on sequence	<a href="http://www.dkfz.de/spec/glycosciences.de/tools/glyseq/">http://www.dkfz.de/spec/glycosciences.de/tools/glyseq/</a>
GlyVicinity [159]	Statistical analysis of glycosylation sites based on 3D structures	<a href="http://www.dkfz.de/spec/glycosciences.de/tools/glyvicinity/">http://www.dkfz.de/spec/glycosciences.de/tools/glyvicinity/</a>

carbohydrates. Quantum mechanics methods are mainly used to study chemical reactions [169–171], to calculate force constants and atom charges to be used as force field parameters [172, 173], or for the conformational analysis of smaller carbohydrates [174–176].

In general, molecular modeling methods are applied in glycobiology at various levels of required expertise and computer equipment. Simple model building using a molecular builder will already give one valuable insight: complex carbohydrates look in 3D very different from the impression one gets by looking at chemical drawings or cartoon representations. Also, the simple overlay in 3D of a new carbohydrate ligand onto an existing one in a crystal structure can give valuable first insight into a possible binding mode. However, one has to be aware of the limitations of such basic modeling approaches: molecular builders generate one reasonable conformation out of many; and manually overlaying two carbohydrates in a binding site is a strong bias towards one predefined binding mode and alternative binding modes are unlikely to be discovered. The other extremes would be to perform a complete conformational analysis of a complex carbohydrate based on extensive MD simulations in explicit solvent, which may take weeks of calculation time, and GBytes of simulation data need to be analyzed afterwards, or to screen the complete protein surface for carbohydrate binding sites using extensive dockings based on genetic search algorithms. Recently, even high-level Car-Parinello-based *ab initio* MD simulations combined with metadynamics simulation have been applied to carbohydrates [170, 177, 178]. The question which modeling method would work best for solving a specific scientific problem is not always straightforward to answer. For example, if exploring the accessible conformational space of a carbohydrate is of major interest then the MD simulation could also be performed in gas phase at higher temperatures instead of running an extensive MD simulation in explicit solvent at room temperature. However, for some systems, the use of explicit solvent is necessary, while for others, one would reach the same conclusions based on gas phase simulations, but with much less computational cost. As is the case with ‘experimental’ methods, ‘experience’ is the key to successful application of molecular modeling methods in most cases.

Since the beginning of the 1990s, more and more crystal structures have been reported where carbohydrates are covalently attached to a (glyco) protein or constitute the ligand in a protein–carbohydrate complex [179, 180]. These experimentally determined 3D structures are freely accessible from the Protein Data Bank (PDB) [112]. So, over recent years, the PDB has become a very valuable resource for obtaining conformational properties of carbohydrates [181]. However, it has to be kept in mind that

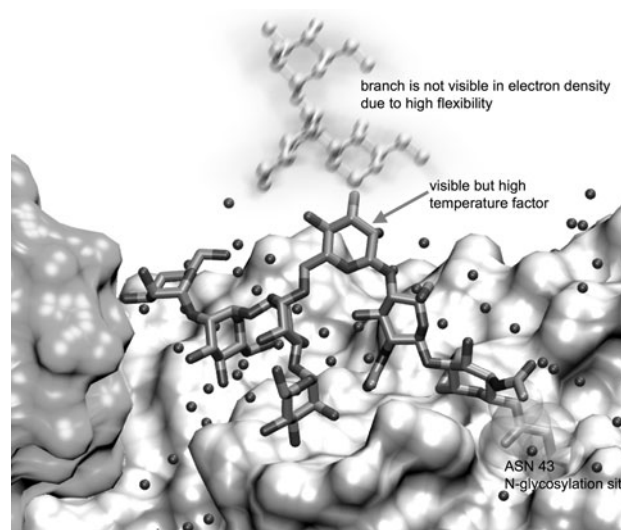
crystals are often grown under non-physiological conditions, and flexible molecules like carbohydrates may change conformation due to strong forces induced by crystallographic packing.

A variety of reviews and book chapters on conformational analysis of carbohydrates have been published and are recommended for further reading [46, 165, 166, 182].

#### Databases containing 3D structures of carbohydrates

The two major databases where experimentally determined carbohydrate structures are stored are the Cambridge Structural Database (CSD, <http://www.ccdc.cam.ac.uk/products/csd/>) and the Protein Data Bank (PDB, <http://www.rcsb.org/pdb/>). Many crystal structures of small oligosaccharides [183] are also accessible through the Glyco3D web interface (<http://www.cermav.cnrs.fr/glyco3d/>).

The PDB [112] currently contains more than 60,000 3D structures of biomolecules, of which about 4,000 contain carbohydrates [184]. Most of the carbohydrates in the PDB are either connected covalently to a (glyco)protein, or the carbohydrate forms a complex with a lectin, enzyme, or antibody. Isolated carbohydrates are only rarely found in the PDB. When looking at the carbohydrate structures of a PDB entry, one has to keep in mind that frequently only fragments of the original carbohydrates may be resolved (Fig. 8). Additionally, the 3D structures of the carbohydrates in the PDB do not always meet high quality standards; therefore, one has to look at the structures with care. It has been recognized that, in order to improve the



**Fig. 8** *N*-Glycosylation site of *Phanerochaete chrysosporium* Laminarinase 16A (pdb code 2W52 [189]). Although the resolution of the X-ray crystal structure is rather high (1.56 Å), parts of the *N*-glycan are not visible in the electron density due to the high flexibility of the branch linked to position 3 of the core mannose

quality of the 3D structures contained in the PDB, theoretical validation procedures for carbohydrates have to be established [184–186]. Despite these limitations, the PDB is an important source of information on carbohydrate 3D structures [187]. Unfortunately, due to the lack of a consistently used nomenclature for carbohydrates in PDB files, it is difficult to find the entries of interest. To overcome this problem, the GLYCOSCIENCES.de web portal [187] and the Glycoconjugate Data Bank: Structures (<http://www.glycostructures.jp>) [188] offer convenient ways to search for carbohydrate structures in the PDB.

Statistical analysis of structural parameters of the carbohydrates present in the PDB entries can be performed with the tools GlySeq, GlyVicinity, GlyTorsion, and carp (<http://www.glycosciences.de/tools>) [159]. GlySeq checks the type of amino acids that are in the sequence neighborhood of *N*- and *O*-glycosylation sites, GlyVicinity performs an analysis of the population of amino acids in the spatial vicinity of carbohydrate residues, and GlyTorsion provides access to the torsion angles of the glycosidic linkages. Using the carp tool (CARbohydrate Ramachandran Plot), these torsions can be compared to theoretical Ramachandran-type conformational maps stored in GlycoMapsDB [185]. This service can be used by crystallographers to cross-check or validate the carbohydrate 3D structures similar to the Ramachandran plot analysis that is routinely used to evaluate the backbone torsions of protein structures.

#### Molecular modeling of carbohydrates over the Internet

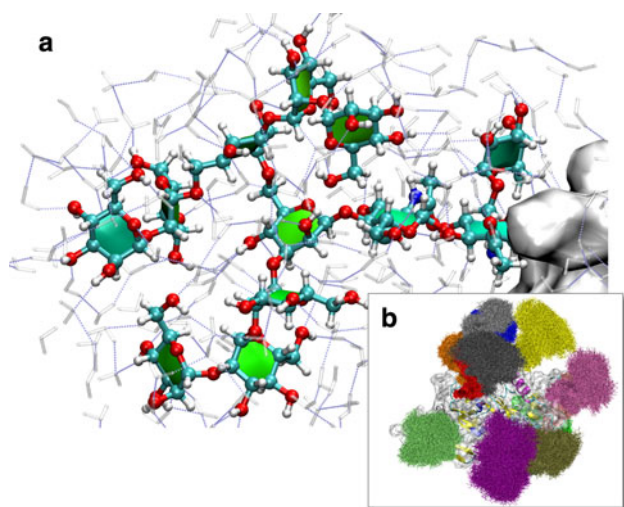
Easy-to-use and freely available Web-based tools [32, 190] are available to generate an initial model of a carbohydrate 3D structure. SWEET-II [191] is a frequently used carbohydrate 3D builder that is available on the GLYCOSCIENCES.de [82] website, which also provides the GlyProt [192] tool for *in silico* glycosylation of proteins derived from the PDB. A very nice builder for carbohydrates and glycoproteins is also available at the GLYCAM website (<http://www.glycam.com>). The first molecular builder for carbohydrates was probably POLYS [193] and the latest development in this field is FSPS (fast sugar structure prediction software) [194]. However, it is unclear whether POLYS or FSPS are available to the scientific community via a website or for download. A web-portal to perform MD simulations of carbohydrates over the Internet [195] was recently shut down because the glycoinformatics group that maintained the service was closed [42].

#### Molecular dynamics simulation

Despite the significant limitations that still exist, the use of molecular dynamics (MD) simulations has turned out to be

an excellent methodology to study the conformational properties of carbohydrates and other biomolecules [163, 196–199]. Although quantum mechanics-based MD simulations have recently become feasible, most applications of MD are still based on force fields. The development of carbohydrate force fields in itself is a challenging task and is still in progress [172, 173, 200]. Because carbohydrates are polar molecules, the proper treatment of atom charges is likely to be of significant importance particularly for modeling of intermolecular interactions [201]. The discussion about including extra terms for (exo) anomeric effects into force fields has a long tradition in carbohydrate modeling [202–205]. The solvent model used for the MD simulation also has a significant effect on the results [206]. In order to make the outcome of an MD simulation more reliable, the theoretical results should always be compared to experimental results if possible. It has to be kept in mind that disagreement between computational and experimental results does not necessarily mean that the force field used is inappropriate for the simulation of carbohydrate structures. It can also mean that other simulation parameters used are not appropriate (e.g., simulation time, solvent model) or possibly that there is a significant error in the experimental results themselves. However, experimental results are very important for validating the quality of theoretical calculations. In recent years, MD simulations have been used to study conformations of complex carbohydrates [206–209], glycolipids [200, 210, 211], glycopeptides [212], glycoproteins [213, 214], protein–carbohydrate complexes [215–217], protein–glycopeptide interaction [218], carbohydrate–ion interaction [219], and carbohydrate–water interaction [220].

The MD simulation of a complex oligosaccharide or glycoprotein in a solvent box is computationally very expensive, and CPU time of many weeks or months may be required in order to simulate a timescale of only a few nanoseconds [221] (Fig. 9). The timescales of most of the published MD simulations involving carbohydrates are in the range of up to 50 ns. However, in order to achieve convergence for the rotamer populations of the exocyclic C–C torsions, the length of an MD simulation should be longer than 100 ns [206, 222]. Although water models like TIP5P are a better approximation of water, in most MD simulations much simpler water models, like SPC or TIP3P [223, 224], are used because of calculation speed and because most force field parameters have been tailored to these simple models. The trajectory files of an MD simulation are typically many GBytes in size. With the availability of supercomputers terabytes of MD data can be produced easily within a short time. The current bottleneck in the application of MD simulations is therefore hard disk space and the requirement to analyze and interpret very quickly the huge amount of data produced. Analysis tools



**Fig. 9** MD simulation of SIV gp120 glycoprotein (M. Frank, unpublished). Complete *N*-glycans were modeled at 13 glycosylation sites based on the X-ray structure (pdb code 2BF1 [229]). **a** Solvation shell of a selected *N*-glycan on the protein surface. **b** A significant surface area of the protein is shielded by the *N*-glycans. The molecular system has more than 100,000 atoms (4,832 protein atoms, 3,432 carbohydrate atoms, 30,665 water molecules, 4 chloride ions). Water molecules are not shown for clarity

that are part of the MD software distributions [225–227], and which have been mainly developed for the analysis of proteins, are frequently used for the analysis of MD trajectories of carbohydrates. However, because of the limitations of the available tools, glycoscientists tend to develop their own ‘in-house’ analysis software. Recently, ‘Conformational Analysis Tools’ (CAT) [228], a novel software for the analysis of MD trajectories, has been made publicly available. CAT is optimized for the efficient conformational analysis of carbohydrates, glycoproteins, and protein–carbohydrate complexes. In summary, MD simulations provide valuable additional information on the conformational dynamics of the system investigated, which is frequently not available from experimental methods. However, although performing simple MD simulations is straightforward in most cases, the correct setup and interpretation of the results requires expert knowledge, and the limitations of the method have to be taken into account.

#### Modeling protein–carbohydrate interaction

One of the major challenges in molecular modeling at the moment is the development of efficient and accurate methods to estimate the binding affinity of protein–carbohydrate complexes [46]. The application of docking methods to study protein–carbohydrate interaction has lately significantly increased [230–236]. Typically, a flexible ligand is docked to a rigid receptor; however, examples are reported where receptor flexibility has been

included in the docking protocol [237–239]. Next to an efficient searching algorithm, the availability of a robust scoring function is critical for the success of docking [240]. Bridging water molecules and CH– $\pi$  interactions [241–243] play a major role in the interaction of a carbohydrate and a protein. However, for various reasons, it is difficult to include these factors into the scoring functions of available standard docking software (e.g. AutoDOCK) [244]. Therefore, despite the many successful applications of docking methods, there are still significant problems with respect to the correct prediction of carbohydrate binding sites and relative affinities in some cases.

If a 3D structure of a protein–carbohydrate complex is available (e.g. an X-ray structure), MD simulations can be performed to study the local interactions (hydrogen bonding, hydrophobic interactions, water bridges) in more detail [245, 246] or to calculate free energy [247–249] and entropy changes upon binding [250]. Since polar OH groups of carbohydrates quite frequently like to bind in areas where water molecules are also found on the protein surface, an investigation of the water binding sites is of particular interest [251–253].

#### Conclusion

From the bioinformatics point of view, carbohydrates are a particularly challenging class of biomolecules. Like nucleic acids or proteins, they are assembled from a set of molecular building blocks; however, due to multiple linkage types and sites, even linear carbohydrates are much more complex. Additionally, complex carbohydrates frequently contain one or more branches, which renders most of the sequence algorithms developed for genes not applicable to carbohydrates, and so more complicated tree-based algorithms have to be developed and applied. As a result, established bioinformatics groups seem to neglect carbohydrates to a large extent, and only a few glycoinformatic pioneers face the challenge to develop computer algorithms for carbohydrate sequences.

Significant improvements in glycan analysis and the application of carbohydrate microarrays in glycomics research have led to a significant increase in the amount of experimental data generated. Unfortunately, because of the lack of an established glycoinformatics infrastructure and standards in the field, each research group or consortium has developed their own storage formats, databases, and tools. This renders data integration and exchange very difficult. In recent years, it has become obvious that this situation needs to be changed in the future, and centrally integrated, curated, and comprehensive databases are required for glycomics, similar to proteomics and genomics [27, 38, 254]. Despite this insight, it is very difficult to



establish a global glycoinformatics infrastructure at the moment due to the lack of funding and leadership. This is particularly unfortunate because, in recent years, the field has made significant progress: a standard glycan sequence format (GlycoCT) has been developed; and GlycomeDB integrates globally all carbohydrate structure databases and makes the structures searchable for scientists through one central web-interface. In the context of the EUROCarbDB project, standards, tools, and databases have been developed to store carbohydrate structures, analytical data, biological context, and literature references. Recently, a database prototype (GlyAffinity) that aims at integrating all types of protein-carbohydrate interaction data has been developed; the bioinformatic cores of the Consortium for Functional Glycomics (CFG) and the Japan Consortium for Glycobiology and Glycotechnology (JCGG) are making available a vast amount of experimental data; and the EuroGlycoArrays Consortium has just entered the field and will provide more experimental data for the community. Last but not least, in recent years, people working in the field have really started to talk to each other, which is an important catalyst for establishing bioinformatics standards and fueling data integration.

Over the years, established methods in structural glycobiochemistry, like X-ray crystallography, NMR, and molecular modeling, have provided valuable insights into the three-dimensional structures of carbohydrates, glycoproteins and protein-carbohydrate complexes. This has further improved our understanding of the functions of glycans, and may help to design better enzymes, drugs, or vaccines [10, 255–258]. It has been realized within the PDB consortium [112] that the data representation and validation of carbohydrates in the PDB needs to be revised. Working groups have been established recently to develop a new format for representing carbohydrates in the PDB, and to recommend new computational tools to be developed [259], as well as to develop standards for glycomics databases and experimental reporting [260].

For a long time glycobiochemistry has been the cinderella field in life sciences: “an area that involves much work but, does not get to show off at the ball with her cousins, the genomes and proteins” [261]. This has changed dramatically over the last 10 years. Large collections of new glycomics data are available that are ready to be integrated into the large data collections of proteomics and genomics. Bioinformatics standards for glycomics have been established and, in the context of the EUROCarbDB project, strategies and concepts for data sharing have been worked out and initial discussions with bioinformatic groups from the proteomics field have taken place. It has been realized that the topics currently discussed in proteomics on data sharing [262] are very similar to the aims of EUROCarbDB. Therefore, the next steps will be to establish a

closer collaboration with bioinformatic groups in proteomics and genomics which will hopefully result in the long-term establishment of glycoinformatic concepts at the European Bioinformatics Institute (EMBL-EBI) or the National Center for Biotechnology Information (NCBI) and a bioinformatics center in Japan. In conclusion, bioinformatics for glycomics has evolved beautifully over recent years and is ready to be invited to show up at the ball with proteomics and genomics in order to waltz together.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

1. Apweiler R, Hermjakob H, Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1473:4–8
2. Marth JD, Grewal PK (2008) Mammalian glycosylation in immunity. *Nat Rev Immunol* 8:874–887
3. Varki A (2008) Sialic acids in human health and disease. *Trends Mol Med* 14:351–360
4. Ohtsubo K, Marth JD (2006) Glycosylation in cellular mechanisms of health and disease. *Cell* 126:855–867
5. Jaeken J, Matthijs G (2007) Congenital disorders of glycosylation: a rapidly expanding disease family. *Annu Rev Genomics Hum Genet* 8:261–278
6. Jefferis R (2009) Recombinant antibody therapeutics: the impact of glycosylation on mechanisms of action. *Trends Pharmacol Sci* 30:356–362
7. Li H, d’Anjou M (2009) Pharmacological significance of glycosylation in therapeutic proteins. *Curr Opin Biotechnol* 20:678–684
8. Kawasaki N, Itoh S, Hashii N, Takakura D, Qin Y, Huang X, Yamaguchi T (2009) The significance of glycosylation analysis in development of biopharmaceuticals. *Biol Pharm Bull* 32:796–800
9. Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA (2007) The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu Rev Immunol* 25:21–50
10. Hecht ML, Stallforth P, Silva DV, Adibekian A, Seeberger PH (2009) Recent advances in carbohydrate-based vaccines. *Curr Opin Chem Biol* 13:354–359
11. Yu U, Lee SH, Kim YJ, Kim S (2004) Bioinformatics in the post-genome era. *J Biochem Mol Biol* 37:75–82
12. Krishnamoorthy L, Mahal LK (2009) Glycomic analysis: an array of technologies. *ACS Chem Biol* 4:715–732
13. Haslam SM, Julien S, Burchell JM, Monk CR, Ceroni A, Garden OA, Dell A (2008) Characterizing the glycome of the mammalian immune system. *Immunol Cell Biol* 86:564–573
14. Zaia J (2008) Mass spectrometry and the emerging field of glycomics. *Chem Biol* 15:881–892
15. Ruhaak LR, Deelder AM, Wührer M (2009) Oligosaccharide analysis by graphitized carbon liquid chromatography-mass spectrometry. *Anal Bioanal Chem* 394:163–174
16. Turnbull JE, Field RA (2007) Emerging glycomics technologies. *Nat Chem Biol* 3:74–77
17. Geyer H, Geyer R (2006) Strategies for analysis of glycoprotein glycosylation. *Biochim Biophys Acta* 1764:1853–1869

18. 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* 376:1–12
19. Karlsson H, Larsson JM, Thomsson KA, Hard I, Backstrom M, Hansson GC (2009) High-throughput and high-sensitivity nano-LC/MS and MS/MS for O-glycan profiling. *Methods Mol Biol* 534:117–131
20. Domann PJ, Pardos-Pardos AC, Fernandes DL, Spencer DI, Radcliffe CM, Royle L, Dwek RA, Rudd PM (2007) Separation-based glycoproteomics approaches using fluorescent labels. *Proteomics* 7(Suppl 1):70–76
21. Wada Y, Azadi P, Costello CE, Dell A, Dwek RA, Geyer H, Geyer R, Kakehi K, Karlsson NG, Kato K, Kawasaki N, Khoo KH, Kim S, Kondo A, Lattova E, Mechref Y, Miyoshi E, Nakamura K, Narimatsu H, Novotny MV, Packer NH, Perreault H, Peter-Katalinic J, Pohlentz G, Reinhold VN, Rudd PM, Suzuki A, Taniguchi N (2007) Comparison of the methods for profiling glycoprotein glycans—HUPO Human Disease Glycomics/Proteome Initiative multi-institutional study. *Glycobiology* 17:411–422
22. Liu Y, Palma AS, Feizi T (2009) Carbohydrate microarrays: key developments in glycobiology. *Biol Chem* 390:647–656
23. Horlacher T, Seeberger PH (2008) Carbohydrate arrays as tools for research and diagnostics. *Chem Soc Rev* 37:1414–1422
24. Hirabayashi J (2008) Concept, strategy and realization of lectin-based glycan profiling. *J Biochem* 144:139–147
25. Pilobello KT, Slawek DE, Mahal LK (2007) A ratiometric lectin microarray approach to analysis of the dynamic mammalian glycome. *Proc Natl Acad Sci USA* 104:11534–11539
26. Raman R, Raguram S, Venkataraman G, Paulson JC, Sasisekharan R (2005) Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat Methods* 2:817–824
27. Packer NH, von der Lieth C-W, Aoki-Kinoshita KF, Lebrilla CB, Paulson JC, Raman R, Rudd P, Sasisekharan R, Taniguchi N, York WS (2008) Frontiers in glycomics: Bioinformatics and biomarkers in disease. An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda, MD (September 11–13, 2006). *Proteomics* 8:8–20
28. Andersson B (2006) European Science Foundation Policy Briefing
29. von der Lieth CW, Lutteke T, Frank M (eds) (2009) Bioinformatics for glycobiology and glycomics: an Introduction. Wiley, New York
30. Kersey P, Apweiler R (2006) Linking publication, gene and protein data. *Nat Cell Biol* 8:1183–1189
31. Mulder NJ, Kersey P, Pruess M, Apweiler R (2008) In silico characterization of proteins: UniProt, InterPro and Integr8. *Mol Biotechnol* 38:165–177
32. Lutteke T (2008) Web Resources for the Glycoscientist. *Chembiochem* 9:2155–2160
33. Mahal LK (2008) Glycomics: towards bioinformatic approaches to understanding glycosylation. *Anticancer Agents Med Chem* 8:37–51
34. Mamitsuka H (2008) Informatic innovations in glycobiology: relevance to drug discovery. *Drug Discov Today* 13:118–123
35. Aoki-Kinoshita KF (2008) An introduction to bioinformatics for glycomics research. *PLoS Comput Biol* 4:e1000075
36. Ranzinger R, Herget S, Lutteke T, Frank M (2009) Carbohydrate Structure Databases. In: Cummings RD, Pierce JM (eds) *Handbook of glycomics*. Elsevier, Amsterdam, pp 211–233
37. von der Lieth C-W (2007) Databases and Informatics for Glycobiology and Glycomics. In: Kamerling JP (ed) *Comprehensive glycoscience—from chemistry to systems biology*, vol 2. Elsevier, Oxford, pp 329–346
38. von der Lieth C-W, Lutteke T, Frank M (2006) The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. *Biochim Biophys Acta* 1760:568–577
39. Aoki-Kinoshita KF, Kanehisa M (2006) Bioinformatics approaches in glycomics and drug discovery. *Curr Opin Mol Ther* 8:514–520
40. Perez S, Mulloy B (2005) Prospects for glycoinformatics. *Curr Opin Struct Biol* 15:517–524
41. Marchal I, Golfier G, Dugas O, Majed M (2003) Bioinformatics in glycobiology. *Biochimie* 85:75–81
42. von der Lieth C-W, Bohne-Lang A, Lohmann KK, Frank M (2004) Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform* 5:164–178
43. Werz DB, Ranzinger R, Herget S, Adibekian A, von der Lieth C-W, Seeberger PH (2007) Exploring the structural diversity of mammalian carbohydrates (“glycospace”) by statistical database analysis. *ACS Chem Biol* 2:685–691
44. Harvey DJ, Merry AH, Royle L, Campbell MP, Dwek RA, Rudd PM (2009) Proposal for a standard system for drawing structural diagrams of N- and O-linked carbohydrates and related compounds. *Proteomics* 9:3796–3801
45. Varki A, Freeze HH, Manzi AE (2009) Overview of glycoconjugate analysis. *Curr Protoc Protein Sci Chapter 12, Unit 12.1* 12.1.1–8
46. DeMarco ML, Woods RJ (2008) Structural glycobiology: a game of snakes and ladders. *Glycobiology* 18:426–440
47. Banin E, Neuberger Y, Altshuler Y, Halevi A, Inbar O, Nir D, Dukler A (2002) A novel Linear Code(R) nomenclature for complex carbohydrates. *TIGG* 14:127–137
48. Herget S, Toukach PV, Ranzinger R, Hull WE, Knirel YA, von der Lieth C-W (2008) Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Struct Biol* 8:35
49. McNaught AD (1997) Nomenclature of carbohydrates (recommendations 1996). *Adv Carbohydr Chem Biochem* 52:43–177
50. 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* 9:5398–5399
51. Ceroni A, Dell A, Haslam SM (2007) The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* 2:3
52. Doubt S, Bock K, Smith D, Darvill A, Albersheim P (1989) The Complex Carbohydrate Structure Database. *Trends Biochem Sci* 14:475–477
53. Pellegrini L, Burke DF, von Delft F, Mulloy B, Blundell TL (2000) Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin. *Nature* 407:1029–1034
54. Murray-Rust P, Mitchell JB, Rzepa HS (2005) Communication and re-use of chemical information in bioscience. *BMC Bioinformatics* 6:180
55. McNaught A (2006) The IUPAC International Chemical Identifier: InChI. *Chemistry International (IUPAC)* 28
56. Weininger D (1988) Smiles, a chemical language and information—system 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28:31–36
57. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 37:W623–W633
58. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 36:D344–D350

59. Herget S, Ranzinger R, Maass K, von der Lieth C-W (2008) GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr Res* 343:2162–2171
60. Ranzinger R, Herget S, Wetter T, von der Lieth C-W (2008) GlycomeDB - integration of open-access carbohydrate structure databases. *BMC Bioinform* 9:384
61. Sahoo SS, Thomas C, Sheth A, Henson C, York WS (2005) GLYDE-an expressive XML standard for the representation of glycan structure. *Carbohydr Res* 340:2802–2807
62. Aoki KF, Yamaguchi A, Ueda N, Akutsu T, Mamitsuka H, Goto S, Kanehisa M (2004) KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res* 32:W267–W272
63. Aoki KF, Mamitsuka H, Akutsu T, Kanehisa M (2005) A score matrix to reveal the hidden links in glycans. *Bioinformatics* 21:1457–1463
64. Hizukuri Y, Yamanishi Y, Nakamura O, Yagi F, Goto S, Kanehisa M (2005) Extraction of leukemia specific glycan motifs in humans by computational glycomics. *Carbohydr Res* 340:2270–2278
65. Kuboyama T, Hirata K, Aoki-Kinoshita KF, Kashima H, Yasuda H (2006) A gram distribution kernel applied to glycan classification and motif extraction. *Genome Inform* 17:25–34
66. Yamanishi Y, Bach F, Vert JP (2007) Glycan classification with tree kernels. *Bioinformatics* 23:1211–1216
67. Hashimoto K, Takigawa I, Shiga M, Kanehisa M, Mamitsuka H (2008) Mining significant tree patterns in carbohydrate sugar chains. *Bioinformatics* 24:i167–i173
68. Rubin DL, Shah NH, Noy NF (2008) Biomedical ontologies: a functional perspective. *Brief Bioinform* 9:75–90
69. Thomas CJ, Sheth A, York WS (2006) In: Proceedings of the International Conference on formal ontology in information systems (FOIS) IOS (in press)
70. York WS, Kochut KJ, Miller JA (2009) Integration of glycomics knowledge and data. In: Cummings RD, Pierce JM (eds) *Handbook of glycomics*. Elsevier, Amsterdam, pp 179–195
71. Laine RA (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields  $1.05 \times 10^{12}$  structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* 4:759–767
72. Umana P, Bailey JE (1997) A mathematical model of N-linked glycoform biosynthesis. *Biotechnol Bioeng* 55:890–908
73. Krambeck FJ, Betenbaugh MJ (2005) A mathematical model of N-linked glycosylation. *Biotechnol Bioeng* 92:711–728
74. Krambeck FJ, Bennun SV, Narang S, Choi S, Yarema KJ, Betenbaugh MJ (2009) A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology* 19:1163–1175
75. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M (2005) Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* 21:3976–3982
76. Suga A, Yamanishi Y, Hashimoto K, Goto S, Kanehisa M (2007) An improved scoring scheme for predicting glycan structures from gene expression data. *Genome Inform* 18:237–246
77. Joshi HJ (2008) Deutsches Krebsforschungszentrum, PhD Thesis, University of Heidelberg
78. Brooksbank C, Camon E, Harris MA, Magrane M, Martin MJ, Mulder N, O'Donovan C, Parkinson H, Tuli MA, Apweiler R, Birney E, Brazma A, Henrick K, Lopez R, Stoesser G, Stoehr P, Cameron G (2003) The European Bioinformatics Institute's data resources. *Nucleic Acids Res* 31:43–50
79. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36:D13–D21
80. Whitfield EJ, Pruess M, Apweiler R (2006) Bioinformatics database infrastructure for biotechnology research. *J Biotechnol* 124:629–639
81. Doubet S, Albersheim P (1992) CarbBank. *Glycobiology* 2:505–507
82. Lutteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, von der Lieth C-W (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* 16:71R–81R
83. 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* 16:63R–70R
84. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R (2006) Advancing glycomics: Implementation strategies at the consortium for functional glycomics. *Glycobiology* 16:82R–90R
85. Toukach FV, Knirel YA (2005) New database of bacterial carbohydrate structures. *Glycoconjugate J* 22:216–217
86. Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* 24:1214–1216
87. Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH (2003) GlycoSuiteDB: a curated relational database of glycoprotein glycan structures, their biological sources. 2003 update. *Nucleic Acids Res* 31:511–513
88. Toukach P, Joshi HJ, Ranzinger R, Knirel Y, von der Lieth C-W (2007) Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de. *Nucleic Acids Res* 35:D280–D286
89. Ranzinger R, Frank M, von der Lieth CW, Herget S (2009) Glycome-DB.org: a portal for querying across the digital world of carbohydrate sequences. *Glycobiology* 19:1563–1567
90. McNaught AD (1997) International Union of Pure and Applied Chemistry and International Union of Biochemistry and Molecular Biology. Joint Commission on Biochemical Nomenclature. Nomenclature of carbohydrates. *Carbohydr Res* 297:1–92
91. Hattori M, Okuno Y, Goto S, Kanehisa M (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* 125:11853–11865
92. Bohne-Lang A, Lang E, Forster T, von der Lieth C-W (2001) LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res* 336:1–11
93. Toukach FV (2009) Bacterial carbohydrate structure database version 3. *Glycoconjugate J* 26:856
94. 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* 29:332–335
95. Maes E, Bonachera F, Strecker G, Guerardel Y (2009) SOACS index: an easy NMR-based query for glycan retrieval. *Carbohydr Res* 344:322–330
96. Shikanai T, Shimma YS, Suzuki YS, Fujita NF, Kaji HK, Sato TS, Togayachi AT, Kameyama AK, Tateno HT, Hirabayashi J, Okuda S, Kawasaki T, Takahashi N, Kato K, Furukawa K, Yasugi E, Nishijima M, Kinoshita K, Nishihara S, Yamada I, Mizuno M, Shirai T, Kato M, Yamaguchi Y, Hagiya E, Yoshida

- K, Taniguchi N, Narimatsu H (2009) Japan consortium for glycobiology and glycochemistry database. *Glycoconjugate J* 26:856–856
97. Kikuchi N, Kameyama A, Nakaya S, Ito H, Sato T, Shikanai T, Takahashi Y, Narimatsu H (2005) The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics* 21:1717–1718
98. Yue T, Haab BB (2009) Microarrays in glycoproteomics research. *Clin Lab Med* 29:15–29
99. Powell AK, Zhi ZL, Turnbull JE (2009) Saccharide microarrays for high-throughput interrogation of glycan-protein binding interactions. *Methods Mol Biol* 534:313–329
100. Culf AS, Cuperlovic-Culf M, Ouellette RJ (2006) Carbohydrate microarrays: survey of fabrication techniques. *Omics* 10:289–310
101. Smith DF, Cummings RD (2009) Glycan-binding proteins and glycan microarrays. In: Cummings RD, Pierce JM (eds) *Handbook of glycomics*. Elsevier, Amsterdam, pp 139–160
102. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, Prieto C, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol* 25:894–898
103. Blixt O, Head S, Mondala T, Scanlan C, Huflejt ME, Alvarez R, Bryan MC, Fazio F, Calarese D, Stevens J, Razi N, Stevens DJ, Skehel JJ, van Die I, Burton DR, Wilson IA, Cummings R, Bovin N, Wong CH, Paulson JC (2004) Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc Natl Acad Sci USA* 101:17033–17038
104. Hirabayashi J (2004) Lectin-based structural glycomics: glycoproteomics and glycan profiling. *Glycoconjugate J* 21:35–40
105. Tateno H, Nakamura-Tsuruta S, Hirabayashi J (2007) Frontal affinity chromatography: sugar-protein interactions. *Nat Protoc* 2:2529–2537
106. Van Damme EJM, Peumans WJ, Pusztai, A, Bardocz S (1998) *Handbook of plant lectins: properties and biomedical applications*, Wiley, Chichester
107. Kilpatrick DC (2000) *Handbook of animal lectins: properties and biomedical applications*. Wiley, Chichester
108. Varki A, Cummings RD, Esko JD, Freeze H, Hart G (ed) (2008) *Essentials of glycobiology*. Cold Spring Harbor Laboratory, New York
109. Porter A, Yue T, Heeringa L, Day S, Suh E, Haab BB (2009) A motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins. *Glycobiology* 20:369–380
110. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* 37:D233–D238
111. Campbell JA, Davies GJ, Bulone V, Henrissat B (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J* 326(Pt 3):929–939
112. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35:D301–D303
113. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res* 37:D588–D592
114. Kikuchi N, Narimatsu H (2006) Bioinformatics for comprehensive finding and analysis of glycosyltransferases. *Biochim Biophys Acta* 1760:578–583
115. Breton C, Snajdrova L, Jeanneau C, Koca J, Imberty A (2006) Structures and mechanisms of glycosyltransferases. *Glycobiology* 16:29R–37R
116. Kawasaki T, Nakao H, Takahashi E, Tominaga T (2006) GlycoEpitope: the integrated database of carbohydrate antigens and antibodies. *TIGG* 18:267–272
117. Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* 27:370–372
118. Kaji H, Kamiie J, Kawakami H, Kido K, Yamauchi Y, Shinkawa T, Taoka M, Takahashi N, Isobe T (2007) Proteomics reveals N-linked glycoprotein diversity in *Caenorhabditis elegans* and suggests an atypical translocation mechanism for integral membrane proteins. *Mol Cell Proteomics* 6:2100–2109
119. Jung E, Veuthey AL, Gasteiger E, Bairoch A (2001) Annotation of glycoproteins in the SWISS-PROT database. *Proteomics* 1:262–268
120. Imberty A, Gerber S, Tran V, Perez S (1990) Data-Bank of 3-Dimensional Structures of Disaccharides, a Tool to Build 3-D Structures of Oligosaccharides. 1. Oligo-Mannose Type N-Glycans. *Glycoconjugate J* 7:27–54
121. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A (2007) UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. *Methods Mol Biol* 406:89–112
122. Gerwig GJ, Vliegthart JF (2000) Analysis of glycoprotein-derived glycopeptides. *Exs* 88:159–186
123. Cooper CA, Gasteiger E, Packer NH (2001) GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* 1:340–349
124. Goldberg D, Sutton-Smith M, Paulson J, Dell A (2005) Automatic annotation of matrix-assisted laser desorption/ionization *N*-glycan spectra. *Proteomics* 5:865–875
125. Goldberg D, Bern M, Li B, Lebrilla CB (2006) Automatic determination of *O*-glycan structure from fragmentation spectra. *J Proteome Res* 5:1429–1434
126. Goldberg D, Bern M, Parry S, Sutton-Smith M, Panico M, Morris HR, Dell A (2007) Automated *N*-glycopeptide identification using a combination of single- and tandem-MS. *J Proteome Res* 6:3995–4005
127. Goldberg D, Bern M, North SJ, Haslam SM, Dell A (2009) Glycan family analysis for deducing *N*-glycan topology from single MS. *Bioinformatics* 25:365–371
128. Gaucher SP, Morrow J, Leary JA (2000) STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal Chem* 72:2331–2336
129. Lapadula AJ, Hatcher PJ, Hanneman AJ, Ashline DJ, Zhang H, Reinhold VN (2005) Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. *Anal Chem* 77:6271–6279
130. Ethier M, Saba JA, Spearman M, Krokhn O, Butler M, Ens W, Standing KG, Perreault H (2003) Application of the StrOligo algorithm for the automated structure assignment of complex *N*-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun Mass Spectrom* 17:2713–2720
131. Tang HX, Mechref Y, Novotny MV (2005) Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* 21:I431–I439
132. Maass K, Ranzinger R, Geyer H, von der Lieth C-W, Geyer R (2007) “Glyco-peakfinder”—de novo composition analysis of glycoconjugates. *Proteomics* 7:4435–4444
133. Lohmann KK, von der Lieth C-W (2004) GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass

- spectra of complex carbohydrates. *Nucleic Acids Res* 32:W261–W266
134. Lohmann KK, von der Lieth C-W (2003) GLYCO-FRAGMENT: A web tool to support the interpretation of mass spectra of complex carbohydrates. *Proteomics* 3:2028–2035
  135. Joshi HJ, Harrison MJ, Schulz BL, Cooper CA, Packer NH, Karlsson NG (2004) Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* 4:1650–1664
  136. 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* 7:1650–1659
  137. Gao HY (2009) Generation of asparagine-linked glycan structure databases and their use. *J Am Soc Mass Spectrom* 20:1739–1742
  138. Ozohanics O, Krenyacz J, Ludanyi K, Pollreis F, Vekey K, Drahos L (2008) GlycoMiner: a new software tool to elucidate glycopeptide composition. *Rapid Commun Mass Spectrom* 22:3245–3254
  139. Vakhrushev SY, Dadimov D, Peter-Katalinic J (2009) Software platform for high-throughput glycomics. *Anal Chem* 81:3252–3260
  140. Tissot B, Ceroni A, Powell AK, Morris HR, Yates EA, Turnbull JE, Gallagher JT, Dell A, Haslam SM (2008) Software tool for the structural determination of glycosaminoglycans by mass spectrometry. *Anal Chem* 80:9204–9212
  141. Clerens S, Van den Ende W, Verhaert P, Geenen L, Arckens L (2004) Sweet Substitute: a software tool for in silico fragmentation of peptide-linked *N*-glycans. *Proteomics* 4:629–632
  142. Irungu J, Go EP, Dalpathado DS, Desaire H (2007) Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID. *Anal Chem* 79:3065–3074
  143. Sturm M, Bertsch A, Gropf C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O (2008) OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinform* 9:163
  144. Sturm M, Kohlbacher O (2009) TOPPView: an open-source viewer for mass spectrometry data. *J Proteome Res* 8:3760–3763
  145. Jansson PE, Kenne L, Widmalm G (1991) Casper—a computer-program used for structural-analysis of carbohydrates. *J Chem Inf Comput Sci* 31:508–516
  146. 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* 341:1003–1010
  147. Fogh RH, Vranken WF, Boucher W, Stevens TJ, Laue ED (2006) A nomenclature and data model to describe NMR experiments. *J Biomol NMR* 36:147–155
  148. Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas P, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins Struct Funct Bioinf* 59:687–696
  149. Julenius K (2007) NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. *Glycobiology* 17:868–876
  150. Gupta R, Brunak S (2002) Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput* pp 310–322
  151. Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S (1998) NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate J*. 15:115–130
  152. Eisenhaber B, Bork P, Yuan Y, Loffler G, Eisenhaber F (2000) Automated annotation of GPI anchor sites: case study *C. elegans*. *Trends Biochem Sci* 25:340–341
  153. Fankhauser N, Maser P (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* 21:1846–1852
  154. Chen YZ, Tang YR, Sheng ZY, Zhang Z (2008) Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of *k*-spaced amino acid pairs. *BMC Bioinform* 9:101
  155. Caragea C, Sinapov J, Silvescu A, Dobbs D, Honavar V (2007) Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinform* 8:438
  156. Hamby SE, Hirst JD (2008) Prediction of glycosylation sites using random forests. *BMC Bioinformatics* 9:500
  157. Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR (2004) Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology* 14:103–114
  158. Thanka Christlet TH, Veluraja K (2001) Database analysis of O-glycosylation sites in proteins. *Biophys J* 80:952–960
  159. Lutteke T, Frank M, von der Lieth C-W (2005) Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PBD. *Nucleic Acids Res* 33:D242–D246
  160. Widmalm G (2007) General NMR spectroscopy of carbohydrates and conformational analysis in solution. In: Kamerling JP (ed) *Comprehensive glycoscience - from chemistry to systems biology*, vol 2. Elsevier, Oxford, pp 101–132
  161. Jimenez-Barbero J, Diaz MD, Nieto PM (2008) NMR structural studies of oligosaccharides related to cancer processes. *Anti-cancer Agents Med Chem* 8:52–63
  162. Lutteke T, Frank M (2009) Synergy of computational and experimental methods in carbohydrate 3D structure determination and validation. In: von der Lieth CW, Lutteke T, Frank M (eds) *Bioinformatics for glycobiology and glycomics: an introduction*. Wiley, New York, pp 389–412
  163. Weimar T, Woods RJ (2003) Combining NMR and simulation methods in oligosaccharide conformational analysis. In: Jimenez-Barbero J, Peters T (eds) *NMR spectroscopy of glycoconjugates*. Wiley, Weinheim, pp 111–144
  164. Frank M (2009) Conformational analysis of carbohydrates—a historical overview. In: von der Lieth CW, Lutteke T, Frank M (eds) *Bioinformatics for glycobiology and glycomics: an introduction*. Wiley, New York, pp 337–357
  165. Frank M (2009) Predicting Carbohydrate 3D Structures Using Theoretical Methods. In: von der Lieth CW, Lutteke T, Frank M (eds.) *Bioinformatics for glycobiology and glycomics: an introduction*. Wiley, New York, pp 359–388
  166. Perez S (2007) Molecular modeling in glycoscience. In: Kamerling JP (ed) *Comprehensive glycoscience - from chemistry to systems biology*, vol 2. Elsevier, Oxford, pp 347–388
  167. Stortz CA (1999) Disaccharide conformational maps: how adiabatic is an adiabatic map? *Carbohydr Res* 322:77–86
  168. von der Lieth C-W, Kozar T, Hull WE (1997) A (critical) survey of modelling protocols used to explore the conformational space of oligosaccharides. *THEOCHEM* 395:225–244
  169. Krupicka M, Tvaroska I (2009) Hybrid quantum mechanical/molecular mechanical investigation of the beta-1, 4-galactosyltransferase-I mechanism. *J Phys Chem B* 113:11314–11319
  170. Dong H, Nimlos MR, Himmel ME, Johnson DK, Qian X (2009) The effects of water on beta-D-xylose condensation reactions. *J Phys Chem A* 113:8577–8585
  171. Greig IR, Zahariev F, Withers SG (2008) Elucidating the nature of the *Streptomyces plicatus* beta-hexosaminidase-bound intermediate using ab initio molecular dynamics simulations. *J Am Chem Soc* 130:17620–17628
  172. 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* 29:622–655
173. Guvench O, Hatcher E, Venable RM, Pastor RW, MacKerell AD (2009) CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J Chem Theory Comput* 5:2353–2370
174. Schnupf U, Willett JL, Bosma W, Momany FA (2009) DFT conformation and energies of amylose fragments at atomic resolution. Part 1: syn forms of alpha-maltotetraose. *Carbohydr Res* 344:362–373
175. Spiwok V, Tvaroska I (2009) Conformational free energy surface of alpha-N-acetylneuraminic acid: an interplay between hydrogen bonding and solvation. *J Phys Chem B* 113:9589–9594
176. Remko M, von der Lieth C-W (2007) Conformational structure of some trimeric and pentameric structural units of heparin. *J Phys Chem A* 111:13484–13491
177. Biarnes X, Ardevol A, Planas A, Rovira C, Laio A, Parrinello M (2007) The conformational free energy landscape of beta-D-glucopyranose. implications for substrate preactivation in beta-glucoside hydrolases. *J Am Chem Soc* 129:10686–10693
178. Laio a, Gervasio FL (2008) Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* 71
179. Qasba PK, Ramakrishnan B (2007) X-ray crystal structures of glycosyltransferases. In: Kamerling JP (ed) *Comprehensive glycoscience—from chemistry to systems biology*, vol 2. Elsevier, Oxford, pp 251–281
180. Buts L, Loris R, Wyns L (2007) X-Ray crystallography of lectins. In: Kamerling JP (ed) *Comprehensive glycoscience - from chemistry to systems biology*, vol 2. Elsevier, Oxford, pp 221–249
181. Lutteke T, von der Lieth CW (2009) Data mining the PDB for glyco-related data. *Methods Mol Biol* 534:293–310
182. Vliegthart JFG, Woods RJ (ed) (2006) *NMR spectroscopy and computer modeling of carbohydrates*. American Chemical Society, Washington, DC
183. Jeffrey GA (1990) Crystallographic studies of carbohydrates. *Acta Crystallogr Sect B Struct Sci* 46(Pt 2):89–103
184. Lutteke T (2009) Analysis and validation of carbohydrate three-dimensional structures. *Acta Crystallogr D Biol Crystallogr* 65:156–168
185. Frank M, Lutteke T, von der Lieth C-W (2007) GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res* 35:287–290
186. Crispin M, Stuart DI, Jones EY (2007) Building meaningful models of glycoproteins. *Nat Struct Mol Biol* 14:354 discussion 354–355
187. Lutteke T, von der Lieth C-W (2006) The protein data bank (PDB) as a versatile resource for glycobiology and glycomics. *Biocatal Biotransform* 24:147–155
188. Nakahara T, Hashimoto R, Nakagawa H, Monde K, Miura N, Nishimura S (2008) Glycoconjugate Data Bank: structures—an annotated glycan structure database and N-glycan primary structure verification service. *Nucleic Acids Res* 36:D368–D371
189. Vasur J, Kawai R, Andersson E, Igarashi K, Sandgren M, Samejima M, Stahlberg J (2009) X-ray crystal structures of *Phanerochaete chrysosporium* Laminarinase 16A in complex with products from lichenin and laminarin hydrolysis. *Febs Journal* 276:4282–4293
190. Berteau O, Stenutz R (2004) Web resources for the carbohydrate chemist. *Carbohydr Res* 339:929–936
191. Bohne A, Lang E, von der Lieth C-W (1999) SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides. *Bioinformatics* 15:767–768
192. Bohne-Lang A, von der Lieth C-W (2005) GlyProt: in silico glycosylation of proteins. *Nucleic Acids Res* 33:W214–W219
193. Engelsen SB, Cros S, Mackie W, Perez S (1996) A molecular builder for carbohydrates: application to polysaccharides and complex carbohydrates. *Biopolymers* 39:417–433
194. Xia J, Margulis C (2008) A tool for the prediction of structures of complex sugars. *J Biomol NMR* 42:241–256
195. Frank M, Gutbrod P, Hassayoun C, von der Lieth C-W (2003) Dynamic molecules: molecular dynamics for everyone. An Internet-based access to molecular dynamic simulations: Basic concepts. *J Mol Model* 9:308–315
196. Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE (2009) Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* 19:120–127
197. Alonso H, Bliznyuk AA, Gready JE (2006) Combining docking and molecular dynamic simulations in drug design. *Med Res Rev* 26:531–568
198. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9:646–652
199. Almond A (2006) Biomolecular dynamics: testing microscopic predictions against macroscopic experiments. In: Vliegthart JFG, Woods RJ (eds.) *NMR spectroscopy and computer modeling of carbohydrates*, vol. 930. American Chemical Society, Washington, DC, pp 156–169
200. Tessier MB, DeMarco ML, Yongye AB, Woods RJ (2008) Extension of the GLYCAM06 biomolecular force field to lipids, lipid bilayers and glycolipids. *Mol Simul* 34:349–363
201. Tschampel SM, Kennerty MR, Woods RJ (2007) TIP5P-consistent treatment of electrostatics for biomolecular simulations. *J Chem Theory Comput* 3:1721–1733
202. Takahashi O, Yamasaki K, Kohno Y, Ueda K, Suezawa H, Nishio M (2009) The origin of the generalized anomeric effect: possibility of CH/n and CH/pi hydrogen bonds. *Carbohydr Res* 344:1225–1229
203. Lii JH, Chen KH, Johnson GP, French AD, Allinger NL (2005) The external-anomeric torsional effect. *Carbohydr Res* 340:853–862
204. Asensio JL, Hidalgo A, Cuesta I, Gonzalez C, Canada J, Vicent C, Chiara JL, Cuevas G, Jimenez-Barbero J (2002) Experimental evidence for the existence of non-exo-anomeric conformations in branched oligosaccharides: NMR analysis of the structure and dynamics of aminoglycosides of the neomycin family. *Chemistry* 8:5228–5240
205. Tvaroska I, Carver JP (1998) The anomeric and exo-anomeric effects of a hydroxyl group and the stereochemistry of the hemiacetal linkage. *Carbohydr Res* 309:1–9
206. Taha HA, Castillo N, Roy PN, Lowary TL (2009) Conformational studies of methyl beta-D-Arabinofuranoside using the AMBER/GLYCAM approach. *J Chem Theory Comput* 5:430–438
207. Olsson U, Sawen E, Stenutz R, Widmalm G (2009) Conformational flexibility and dynamics of two (1->6)-linked disaccharides related to an oligosaccharide epitope expressed on malignant tumour cells. *Chemistry* 15:8886–8894
208. Sanchez-Medina I, Frank M, von der Lieth CW, Kamerling JP (2009) Conformational analysis of the neutral exopolysaccharide produced by *Lactobacillus delbrueckii* ssp. *bulgaricus* LBB.B26. *Org Biomol Chem* 7:280–287
209. Yongye AB, Gonzalez-Outeirino J, Glushka J, Schultheis V, Woods RJ (2008) The conformational properties of methyl alpha-(2, 8)-Di/Trisialosides and their N-Acyl analogues: implications for anti-neisseria meningitidis B vaccine design. *Biochemistry* 47:12493–12514
210. Jedlovsky P, Sega M, Vallauri R (2009) GM1 ganglioside embedded in a hydrated DOPC membrane: a molecular dynamics simulation study. *J Phys Chem B* 113:4876–4886

211. DeMarco ML, Woods RJ (2009) Atomic-resolution conformational analysis of the GM3 ganglioside in a lipid bilayer and its implications for ganglioside-protein recognition at membrane surfaces. *Glycobiology* 19:344–355
212. Fernandez-Tejada A, Corzana F, Busto JH, Jimenez-Oses G, Jimenez-Barbero J, Avenoza A, Peregrina JM (2009) Insights into the geometrical features underlying beta-*O*-GlcNAc glycosylation: water pockets drastically modulate the interactions between the carbohydrate and the peptide backbone. *Chemistry* 15:7297–7301
213. Blanchard V, Frank M, Leefflang BR, Boelens R, Kamerling JP (2008) The structural basis of the difference in sensitivity for PNGase F in the de-N-glycosylation of the native bovine pancreatic ribonucleases B and BS. *Biochemistry* 47:3435–3446
214. Choi Y, Lee JH, Hwang S, Kim JK, Jeong K, Jung S (2008) Retardation of the unfolding process by single N-glycosylation of ribonuclease A based on molecular dynamics simulations. *Biopolymers* 89:114–123
215. Xu D, Newhouse EI, Amaro RE, Pao HC, Cheng LS, Markwick PR, McCammon JA, Li WW, Arzberger PW (2009) Distinct glycan topology for avian and human sialopentasaccharide receptor analogues upon binding different hemagglutinins: a molecular dynamics perspective. *J Mol Biol* 387:465–491
216. Mackeen MM, Almond A, Deschamps M, Cumpsteys I, Fairbanks AJ, Tsang C, Rudd PM, Butters TD, Dwek RA, Wormald MR (2009) The conformational properties of the Glc3Man unit suggest conformational biasing within the chaperone-assisted glycoprotein folding pathway. *J Mol Biol* 387:335–347
217. Mark P, Baumann MJ, Eklof JM, Gullfot F, Michel G, Kallas AM, Teeri TT, Brumer H, Czjzek M (2009) Analysis of nasturtium TmNXG1 complexes by crystallography and molecular dynamics provides detailed insight into substrate recognition by family GH16 xyloglucan endo-transglycosylases and endohydrolases. *Proteins* 75:820–836
218. Gunnerson KN, Pereverzev YV, Prezhdo OV (2009) Atomistic simulation combined with analytic theory to study the response of the P-selectin/PSGL-1 complex to an external force. *J Phys Chem B* 113:2090–2100
219. Eriksson M, Lindhorst TK, Hartke B (2008) Differential effects of oligosaccharides on the hydration of simple cations. *J Chem Phys* 128:105105
220. Ramadugu SK, Chung YH, Xia JC, Margulis CJ (2009) When sugars get wet. A comprehensive study of the behavior of water on the surface of oligosaccharides. *J Phys Chem B* 113:11003–11015
221. Nadas J, Li C, Wang PG (2009) Computational structure activity relationship studies on the CD1d/glycolipid/TCR complex using AMBER and AUTODOCK. *J Chem Inf Model* 49:410–423
222. Gonzalez-Outeirino J, Kirschner KN, Thobhani S, Woods RJ (2006) Reconciling solvent effects on rotamer populations in carbohydrates—a joint MD and NMR analysis. *Can J Chem* 84:569–579
223. Jorgensen WL, Tirado-Rives J (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci USA* 102:6665–6670
224. Dill KA, Truskett TM, Vlachy V, Hribar-Lee B (2005) Modeling water, the hydrophobic effect, and ion solvation. *Ann Rev Biophys Biomol Struct* 34:173–199
225. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ochinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. *J Comput Chem* 30:1545–1614
226. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26:1668–1688
227. Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) Gromacs: fast, flexible, and free. *J Comput Chem* 26:1701–1718
228. Conformational Analysis Tools (CAT), URL: <http://www.md-simulations.de/CAT/>
229. Chen B, Vogan EM, Gong H, Skehel JJ, Wiley DC, Harrison SC (2005) Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature* 433:834–841
230. Chen Y, Shoichet BK (2009) Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat Chem Biol* 5:358–364
231. Reina JJ, Diaz I, Nieto PM, Campillo NE, Paez JA, Tabarani G, Fieschi F, Rojo J (2008) Docking, synthesis, and NMR studies of mannosyl trisaccharide ligands for DC-SIGN lectin. *Org Biomol Chem* 6:2743–2754
232. Takaoka T, Mori K, Okimoto N, Neya S, Hoshino T (2007) Prediction of the structure of complexes comprised of proteins and glycosaminoglycans using docking simulation and cluster analysis. *J Chem Theory Comput* 3:2347–2356
233. Nurisso A, Kozmon S, Imberty A (2008) Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the carbohydrate binding mode to sea cucumber lectin CEL-III. *Molecular Simulation* 34:469–479
234. Guerrini M, Guglieri S, Casu B, Torri G, Mourier P, Boudier C, Viskov C (2008) Antithrombin-binding octasaccharides and role of extensions of the active pentasaccharide sequence in the specificity and strength of interaction. Evidence for very high affinity induced by an unusual glucuronic acid residue. *J Biol Chem* 283:26662–26675
235. de Geus DC, van Roon AMM, Thomassen EAJ, Hokke CH, Deelder AM, Abrahams JP (2009) Characterization of a diagnostic Fab fragment binding trimeric Lewis X. *Proteins Struct Funct Bioinform* 76:439–447
236. Laederach A, Reilly PJ (2005) Modeling protein recognition of carbohydrates. *Proteins* 60:591–597
237. Voss C, Eyol E, Frank M, von der Lieth C-W, Berger MR (2006) Identification and characterization of riproximin, a new type II ribosome-inactivating protein with antineoplastic activity from *Ximenia americana*. *FASEB J* 20:1194–1196
238. Moitessier N, Westhof E, Hanessian S (2006) Docking of aminoglycosides to hydrated and flexible RNA. *J Med Chem* 49:1023–1033
239. Landon MR, Amaro RE, Baron R, Ngan CH, Ozonoff D, McCammon JA, Vajda S (2008) Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. *Chem Biol Drug Des* 71:106–116
240. Hill AD, Reilly PJ (2008) A Gibbs free energy correlation for automated docking of carbohydrates. *J Comput Chem* 29:1131–1141
241. Raju RK, Ramraj A, Hillier IH, Vincent MA, Burton NA (2009) Carbohydrate-aromatic pi interactions: a test of density functionals and the DFT-D method. *Phys Chem Chem Phys* 11:3411–3416
242. Spiwok V, Lipovova P, Skalova T, Vondrackova E, Dohnalek J, Hasek J, Kralova B (2005) Modelling of carbohydrate-aromatic interactions: ab initio energetics and force field performance. *J Comput Aided Mol Des* 19:887–901
243. Vandenburg S, Diaz D, Fernandez-Alonso MC, Pan W, Vincent SP, Cuevas G, Canada FJ, Jimenez-Barbero J, Bartik K (2008) Aromatic-carbohydrate interactions: an NMR and computational study of model systems. *Chemistry* 14:7570–7578

244. Kerzmann A, Fuhrmann J, Kohlbacher O, Neumann D (2008) BALLDock/SLICK: A new method for protein-carbohydrate docking. *Journal of Chemical Information and Modeling* 48:1616–1625
245. Meynier C, Guerlesquin F, Roche P (2009) Computational studies of human galectin-1: role of conserved tryptophan residue in stacking interaction with carbohydrate ligands. *J Biomol Struct Dyn* 27:49–58
246. Mishra NK, Kulhanek P, Snajdrova L, Petrek M, Imberty A, Koca J (2008) Molecular dynamics study of *Pseudomonas aeruginosa* lectin-II complexed with monosaccharides. *Proteins* 72:382–392
247. Fujimoto YK, Terbush RN, Patsalo V, Green DF (2008) Computational models explain the oligosaccharide specificity of cyanovirin-N. *Protein Sci* 17:2008–2014
248. Das P, Li JY, Royyuru AK, Zhou RH (2009) Free energy simulations reveal a double mutant avian H5N1 virus hemagglutinin with altered receptor binding specificity. *J Comput Chem* 30:1654–1663
249. Cai W, Sun T, Liu P, Chipot C, Shao X (2009) Inclusion mechanism of steroid drugs into beta-cyclodextrins. Insights from free energy calculations. *J Phys Chem B* 113:7836–7843
250. Diehl C, Genheden S, Modig K, Ryde U, Akke M (2009) Conformational entropy changes upon lactose binding to the carbohydrate recognition domain of galectin-3. *J Biomol NMR* 45:157–169
251. Gauto DF, Di Lella S, Guardia CM, Estrin DA, Marti MA (2009) Carbohydrate-binding proteins: Dissecting ligand structures through solvent environment occupancy. *J Phys Chem B* 113:8717–8724
252. Di Lella S, Ma L, Ricci JC, Rabinovich GA, Asher SA, Alvarez RM (2009) Critical role of the solvent environment in galectin-1 binding to the disaccharide lactose. *Biochemistry* 48:786–791
253. Kadirvelraj R, Foley BL, Dyekjaer JD, Woods RJ (2008) Involvement of water in carbohydrate-protein binding: concanavalin A revisited. *J Am Chem Soc* 130:16933–16942
254. Seeberger PH (2009) Chemical glycobiology: why now? *Nat Chem Biol* 5:368–372
255. Shaikh FA, Withers SG (2008) Teaching old enzymes new tricks: engineering and evolution of glycosidases and glycosyl transferases for improved glycoside synthesis. *Biochem Cell Biol* 86:169–177
256. Vliegthart JF (2006) Carbohydrate based vaccines. *FEBS Lett.* 580:2945–2950
257. Debeljak N, Sytkowski AJ (2008) Erythropoietin: new approaches to improved molecular designs and therapeutic alternatives. *Curr Pharm Des* 14:1302–1310
258. von Itzstein M, Thomson R (2009) Anti-influenza drugs: the development of sialidase inhibitors. *Handb Exp Pharmacol* 111–154
259. CFG Workshop on Leveraging Glycan Array Screens with Biological, Computational and structural data. <http://glycomics.scripps.edu/CFGWorkshopOct2009.html>
260. CFG Workshop on Analytic and Bioinformatic Glycomics, <http://glycomics.scripps.edu/CFGWorkshopApril2009.html>
261. Hurlley S, Service R, Szuromi P (2001) Cinderella's coach is ready. *Science* 291:2337–2337
262. Rodriguez H, Snyder M, Uhlen M, Andrews P, Beavis R, Borchers C, Chalkley RJ, Cho SY, Cottingham K, Dunn M, Dylag T, Edgar R, Hare P, Heck AJ, Hirsch RF, Kennedy K, Kolar P, Kraus HJ, Mallick P, Nesvizhskii A, Ping P, Ponten F, Yang L, Yates JR, Stein SE, Hermjakob H, Kinsinger CR, Apweiler R (2009) Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principles. *J Proteome Res* 8:3689–3692