Glycobiology vol. 16 no. 5 pp. 71R–81R, 2006
doi:10.1093/glycob/cwj049
Advance Access publication on October 20, 2005

GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research

Thomas Lütteke[2], Andreas Bohne-Lang[3], Alexander Loss, Thomas Goetz, Martin Frank[4], and Claus-W. von der Lieth[1]

Spectroscopic Department (B090), German Cancer Research Center, Molecular Modelling, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Received on August 25, 2005; revised on October 7, 2005; accepted on October 11, 2005

The development of glycan-related databases and bioinformatics applications is considerably lagging behind compared with the wealth of available data and software tools in genomics and proteomics. Because the encoding of glycan structures is more complex, most of the bioinformatics approaches cannot be applied to glycan structures. No standard procedures exist where glycan structures found in various species, organs, tissues or cells can be routinely deposited. In this article the concepts of the GLYCOSCIENCES.de portal are described. It is demonstrated how an efficient structure-based cross-linking of various glycan-related data originating from different resources can be accomplished using a single user interface. The structure oriented retrieval options—exact structure, substructure, motif, composition and sugar components—are discussed. The types of available data—references, composition, spatial structures, nuclear magnetic resonance (NMR) shifts (experimental and estimated), theoretically calculated fragments and Protein Database (PDB) entries—are exemplified for Man$_3$. The free availability and unrestricted use of glycan-related data is an absolute prerequisite to efficiently share distributed resources. Additionally, there is an urgent need to agree to a generally accepted exchange format as well as to a common software interface. An open access repository for glyco-related experimental data will secure that the loss of primary data will be considerably reduced.

Key words: databases/glycobioinformatics/glycomics/structure retrieval

Introduction

With the awareness that the human genome encodes for a significant smaller number of genes than estimated from genomes of lower organisms like yeast (Sequencing, 2004), it became obvious that each gene can be used in different ways depending on how it is regulated. Consequently, the study of posttranslational protein modifications, which can alter the functions of proteins, came increasingly into the scientific focus. Glycobiology research has attracted increasing attention because glycosylation is the most complex and most frequently occurring posttranslational modification (Dell and Morris, 2001). Similar to the development in genomics and proteomics, high throughput glycomics projects to decipher the role of carbohydrates in health and disease are emerging (Blixt *et al.*, 2004; Feizi and Chai, 2004; Pratt and Bertozzi, 2005; Wong, 2005). With the increasing amount of experimental data the need to develop appropriate glycan-related databases and bioinformatics tools is obvious, however, until recently informatics have been poorly involved in glycobiology (von der Lieth *et al.*, 2004). Because the encoding of glycan structures is more complex than those of DNA, RNA and protein sequences, most of the bioinformatics approaches like similarity searches cannot be directly applied to glycan structures. Additionally, no standard procedures exist where scientists can routinely deposit the glycan structures they found in various species, organs, tissues or cells.

Such data are spread over many publications, using different formats and representations. Because a variety of graphical cartoons are often used to report the detected glycan structures, the retrospective extraction of these data has to be done by well-trained people, which is time intensive and costly. Consequently, no comprehensive routinely updated freely available database for glycans is currently available.

The Complex Carbohydrate Structure Database (CCSD) (Doubet *et al.*, 1989; Doubet and Albersheim, 1992)—often named as *CarbBank* according to the retrieval software to access the data—was developed and maintained by the Complex Carbohydrate Research Centre of the University of Georgia (USA). It was the largest effort during the 1990s to collect glycan structures mainly through retrospective manual extraction from literature. The main issue of this approach was to easily find all publications in which specific carbohydrate structures were reported. However, when the funding stopped during the second half of the 1990s, *CarbBank* was not further developed and CCSD no longer updated. Nevertheless, with about 45,000 entries of about 20,000 different structures, the CCSD is still the largest repository of glycan-related data. The text-oriented structural description of glycans as developed by *CarbBank* has been adopted by most of the newer approaches. Also the three larger database projects which emerged in recent years—Kyoto Encyclopedia of Genes and Genomes (KEGG)-glycan (Aoki *et al.*, 2004; Hashimoto *et al.*, 2005), the bioinformatics core of the US Consortium for Functional Glycomics (CFG) (http://www.functionalglycomics.org) as well as the GLYCOSCIENCES.de portal (http://www.glycosciences.de) (Goetz *et al.*, 2004) [the former

[1] To whom correspondence should be addressed; e-mail:w.vonderlieth @dkfz.de

[2] Present address: Division of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139

[3] Present address: Mannheim Medical Research Center, University of Heidelberg, Theodor-Kutzer-Ufer 1-3, D-68135 Mannheim, Germany

[4] Present address: Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, NL-3584 CH Utrecht, The Netherlands

SWEET-DB (Loss *et al.*, 2002)]—provide access to CCSD entries and enable amongst others a *CarbBank*-like (Doubet *et al.*, 1989; Doubet and Albersheim, 1992) representation of glycan structures. However, depending on the scientific background each of the three groups originally came from, the CCSD glycan data have been incorporated in diverse ways to support research in different areas of glycobiology. All three implementations in common are that the retrieval of glycan structures and related data is offered through a (sub)structure search. A speciality of the KEGG-glycan approach (Aoki *et al.*, 2003, 2005) is to provide access to sugar structures through manually drawn pathway maps representing current knowledge on glycan biosynthesis and metabolism of various species. As all known enzymes involved in making and degrading the glycans are also reported, a direct link to genomics and proteomics data is established. The CFG databases are primarily designed to include experimental data produced by the various cores of the consortium. An especially useful approach is the glycan array screening facility. It is the first freely available database presenting reliable and consistent data on protein carbohydrate interactions. In such a way access to carbohydrates which are recognized by a specific lectin are provided. The GLYCOSCIENCES.de portal is an attempt to link glycan-related data originating from various resources through a unique structural description (see supplementary material 1). Special emphasis has been made to provide an easy access to the available experimentally determined spatial structures of glycans and to analyze their interactions with proteins.

In this article, we will present the underlying concepts of the GLYCOSCIENCES.de portal and describe its current status with emphasis on recently added new features. Additionally, we will demonstrate how an efficient structure-based cross-linking of various glycan-related data originating from different resources can be accomplished using a single-user interface. Finally, we will discuss the required next steps to link smaller projects into a powerful network of online connected glycan databases.

Basic concepts

Comparable to data collections of gene and protein sequences, glycan structures are a logical element to arrange all associated data including experiments, references, spectra and biological occurrence. However, since oligosaccharides can exhibit varying and multiple linkages, they cannot be described as simple sequences. To use a glycan structure as identifier in database applications, a special encoding scheme is required, which transforms branched oligosaccharide chains into a unique linear description. The GLYCOSCIENCES.de portal uses the LInear Notation for Unique description of Carbohydrate Sequences (LINUCS) notation (Bohne-Lang *et al.*, 2001) to uniquely encode fully characterized glycans. Structurally fully characterized are those glycans, where all linkage position as well as the stereochemistry of all residues are unambiguously defined. The structure-oriented approach has the advantage that data of the same glycan originating from various sources—like nuclear magnetic resonance (NMR) and mass spectrometry (MS) spectra—can be easily linked and accessed using a single database query.

Another practical complicacy to develop informatics tools for glycosciences originates from the fact that different scientific communities use different ways to represent complex carbohydrate structures. Medically and biologically oriented scientists prefer to report compositions or cartoon-like representation. In contrast chemists favor a complete structural description either in form of an implicit identification of the stereochemistry through symbolic names (e.g., a-D-Galp) or through the use of pseudo three-dimensional (3D) plots, where the stereochemistry of each atom is explicitly drawn utilizing wedges to indicate the spatial orientations of bonds. On the other hand X-ray crystallographers, NMR spectroscopists and molecular modelers often prefer 3D representations of glycans. Although all mentioned representations of glycan structures should be offered and users should have a choice to select their preferred depiction, only one internal description—this is the LINUCS notation in GLYCOSCIENCES.de—is in principle required from which all the other representations can be automatically generated.

There is currently an ongoing debate about the most convenient way how users can input carbohydrate structures. Java-based graphical approaches (Kikuchi *et al.*, 2005) are discussed, as well as tools where monosaccharide residues and their linkages can be selected from pull-down menus and input into spreadsheets (Cooper *et al.*, 2003). The GLYCOSCIENCES.de portal currently supports only the input using the so called text-oriented extended IUPAC description, which was introduced by *CarbBank*. This approach has the advantage, that a carbohydrate structure can be easily transferred between various applications using the normal copy–paste mechanism provided by any browser. Because graphically oriented input facilities are now freely available like KEGG-Draw (Aoki *et al.*, 2004), it is aimed to integrate such tools. For novice users and those who only seldom use the portal, an intuitive graphical interface will reduce the barrier to access the data.

The glycan structure may not be the only desired way to retrieve data. Standard database technology allows storing data in separate tables, which can be cross-referenced by unique identifies. In principle, any stored data can be used for user queries, provided the underlying data model is appropriately structured according to the needs of specific scientific questions. Besides various options of structural retrieval the GLYCOSCIENCES.de portal has implemented options to access glycan-related data through the input of experimental data like NMR, MS, crystallographic data or biological occurrence.

Structural retrieval

According to the varying needs of specific research questions, the GLYCOSCIENCES.de portal provides several structure-oriented options to recall glycan-related data. The retrieval of exactly matching glycans is the most traditional way to access a database. Here, it is accomplished through the lookup of the LINUCS (Bohne-Lang *et al.*, 2001) notation, entered directly or translated from a IUPAC representation entered by the user (see supplementary material 2). Since entries in the database often have identical carbohydrates but varying attached nonsugar components, an option has been implemented allowing to retrieve sugar components where varying aglyca are ignored (see supplementary material 3) and all entries with uncertain linkages which match the fully assigned structure are displayed (see supplementary material 4).

To retrieve glycans based on composition of residues is an especially useful option for larger structures, which also exhibit residues whose masses differ from that of frequently occurring hexoses (see supplementary material 5/6). It can be searched for the exact number of contained residues as well as ranges (e.g., between 2 and 6 Hex or more than 3 NeuAc residues).

Substructure search

Substructure search is the most frequently used way to look for glycan structures. The GLYCOSCIENCES.de portal offers a so-called “beginner” spreadsheet, where a limited number of residues and linkages can be selected from a pull-down menu. The advanced substructure search requires to manually type the name of to be retrieved monosaccharide units and their linkages. Wildcards either for a single character (?)—e.g., if it is unknown if the anomeric carbon is α or β—or for any zero or more characters (*) are supported. Figure 1 depicts the search query for a trisaccharide substructure with various wildcards. Some typical retrieved structures are also shown where the matched substructures are indicated in red. The advanced mode allows to search for two substructures simultaneously, which can be logically connected by an “AND” or an “OR” operation. Figure 2 shows some examples of structures, which contain the *N*-glycan core as well as the Lewis^{X} motif. For both substructure retrieval options, the user can indicate whether he/she wants to access only entries, for which specific resources like NMR spectra or PDB structures are available. For all retrieved structures, direct links to the other stored data are enabled activating the corresponding buttons.

Two substructures, each containing up to five residues, can be searched in a single query. Because typical carbohydrate structures exhibit between three and about twelve residues, the implemented search is capable to deal with most types of substructures.

Motif search

The motif search enables to retrieve all entries, which exhibit substructures having names like Lewis^{X}, blood group H antigen or GM3 (see supplementary material 7). Currently, about 50 frequently used motifs are retrievable. This list can be easily expanded, since simply the connections of residues for a new substructure, its representation using a IUPAC text description and the associated common name have to be added to an editable ASCII-file. The motif search also allows searching for *O*-glycan core structures (Figures 3 and 4).

Because *N*-glycans show a structural high complexity, a separate search interface was implemented (see supplementary material 8) to customize the query by looking for types of *N*-glycans (high mannose, complex or hybrid), the number of antennas, the type and the number of terminating residues as well as bisecting and core-fucosylation. Additionally, the retrieval of *N*-glycans can be combined with the motif search (see Figure 4 for some typical results).

Linked databases

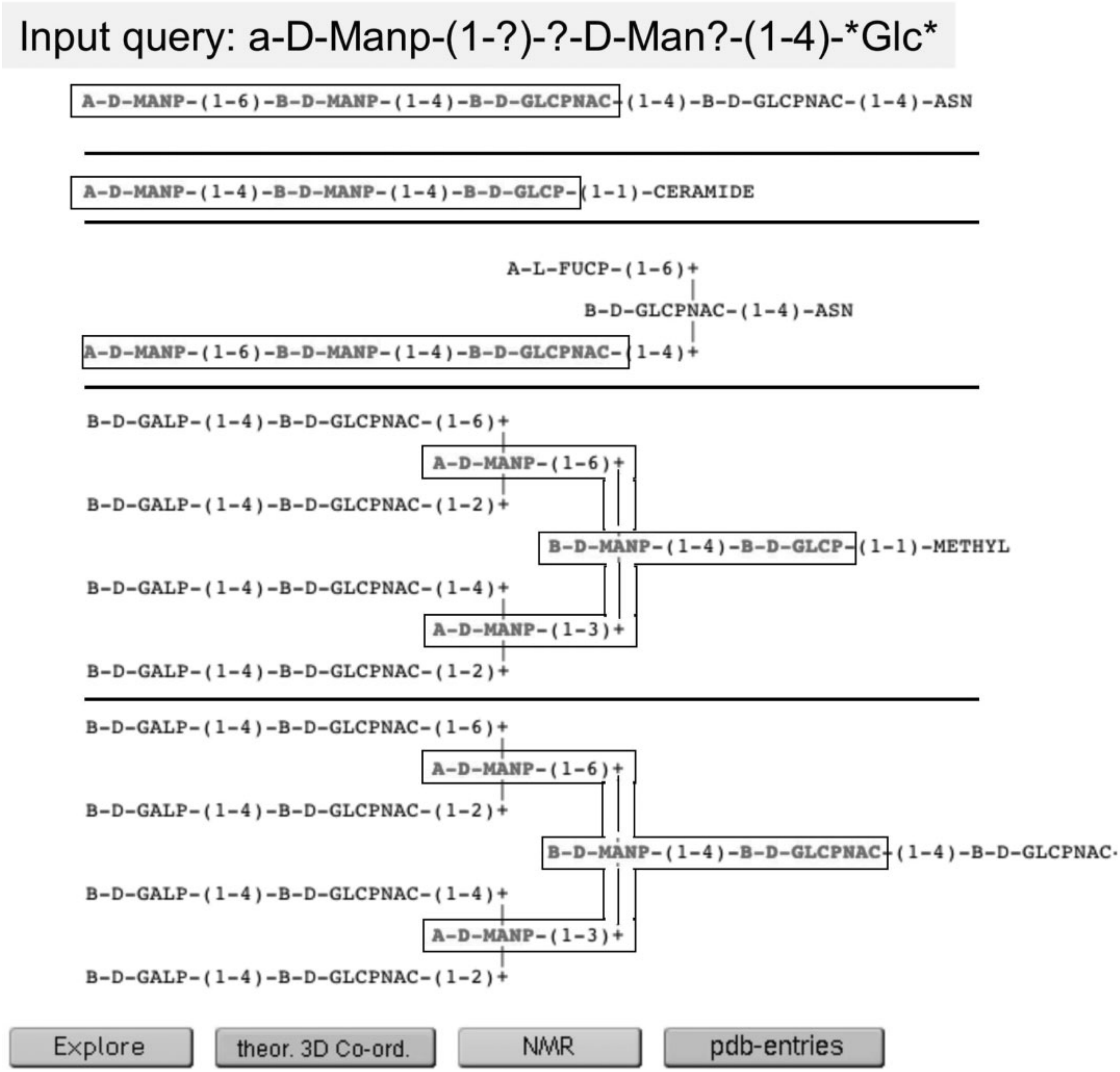
The various structure-based retrieval options provide access to several data resources:

1. References taken from CCSD up to 1997 and some newly entered references.
2. 3D coordinates automatically generated with the *SWEET-II* service (Bohne *et al.*, 1999).
3. ^{1}H and ^{13}C NMR-spectra: about 1500 spectra taken from Sugabase (van Kuik and Vliegenthart, 1992; van Kuik *et al.*, 1992) plus about 2000 NMR spectra manually extracted from the literature, both represented as lists of peaks, which are assigned to a certain atom.
4. Estimated ^{1}H and ^{13}C NMR-shift lists for those structures, where no experimental NMR spectra are available.
5. Masses of theoretically calculated fragments (Lohmann and von der Lieth, 2003, 2004).
6. Ligands, *N*- and *O*-glycans contained in PDB. The automatic assignment of carbohydrate chains is based on 3D coordinates and their connectivity (Lütteke *et al.*, 2004).
7. 3D conformational maps for many glycosidic linkages.

Figure 5 depicts the type of experimental, thereof derived and other generated data, which can be retrieved through the GLYCOSCIENCES.de portal for a specific carbohydrate. Here the *N*-glycan core region (Man$_{3}$) is taken as an example as provided for the exact structure search (click on Trimannsoyl core *N*-glycan to activate Man$_{3}$). At first, the user gets an overview where the availability of each type of data is indicated. Activating the corresponding fields a full display of the associated data is presented. By default only the carbohydrate structure will be displayed using the *CarbBank* representation. In Figure 5a the contained motif, the general structure data as well as the composition are depicted.

Theoretical 3D structures

Clicking on the “theor(etical) 3D co-ord(inate)” button below the structure, a spatial structure of Man$_{3}$ is displayed, which has been automatically generated using the *SWEET-II* service (Bohne *et al.*, 1999). Subsequently the initially constructed geometry was optimized with the *MM3* force field as implemented in die *TINKER*-software package (http://dasher.wustl.edu/tinker/). The *JMOL* (http://jmol.sourceforge.net/) *JAVA* applet is used for the 3D molecular display. Because *JMOL* performs well for all commonly used browsers, no additional software has to be installed locally by the user. One can easily rotate the molecule, choose between various display options and input *RASMOL* (Sayle and Milner-White, 1995) script commands, for example, to color-code specific parts of the molecule. The generated coordinates can be saved and used as


Input query: a-D-Manp-(1-?)-?-D-Man?-(1-4)-*Glc*

```
A-D-MANP-(1-6)-B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)-B-D-GLCPNAC-(1-4)-ASN

A-D-MANP-(1-4)-B-D-MANP-(1-4)-B-D-GLCP-(1-1)-CERAMIDE

                                  A-L-FUCP-(1-6)+
                                                |
                                      B-D-GLCPNAC-(1-4)-ASN
                                                |
A-D-MANP-(1-6)-B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)+

B-D-GALP-(1-4)-B-D-GLCPNAC-(1-6)+
                                |
                  A-D-MANP-(1-6)+
                                |
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)+
                                 B-D-MANP-(1-4)-B-D-GLCP-(1-1)-METHYL
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-4)+
                                |
                  A-D-MANP-(1-3)+
                                |
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)+

B-D-GALP-(1-4)-B-D-GLCPNAC-(1-6)+
                                |
                  A-D-MANP-(1-6)+
                                |
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)+
                                 B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)-B-D-GLCPNAC
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-4)+
                                |
                  A-D-MANP-(1-3)+
                                |
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)+
```

Fig. 1. Example for a substructure search using wildcards. Two types of wildcards are supported: (?) for a single character and (*) for any zero or more characters. Four typical examples are shown which match the query displayed at the top. The found substructures are indicated with a box. For each retrieved entry, the available associated data can be displayed by activating the corresponding button.

input for other computational approaches like molecular dynamics simulations or docking procedures.

NMR

If measured ^{1}H and ^{13}C-NMR shifts are available, the stored data will be presented when activating the corresponding button (Figure 5c). A list of all shifts and their assignments to atoms is displayed. The linkage path—this is the list of glycosidic attachment positions starting from the non-reducing end—is used to provide a unique identification of each residue within a given glycan chain. If no—or not for all atoms of a given structure—experimental shifts are stored, a procedure to estimate ^{1}H and ^{13}C-NMR is implemented (Figure 5d). The procedure to estimate NMR shifts can be invoked by following the NMR link on the databases tab at the top of the page. The NMR shift estimation is based on an appropriate encoding of the structural environment for each atom in the database, which is stored in a table together with the assigned shift value. For the estimation of ^{1}H and ^{13}C-NMR shifts, the same encoding

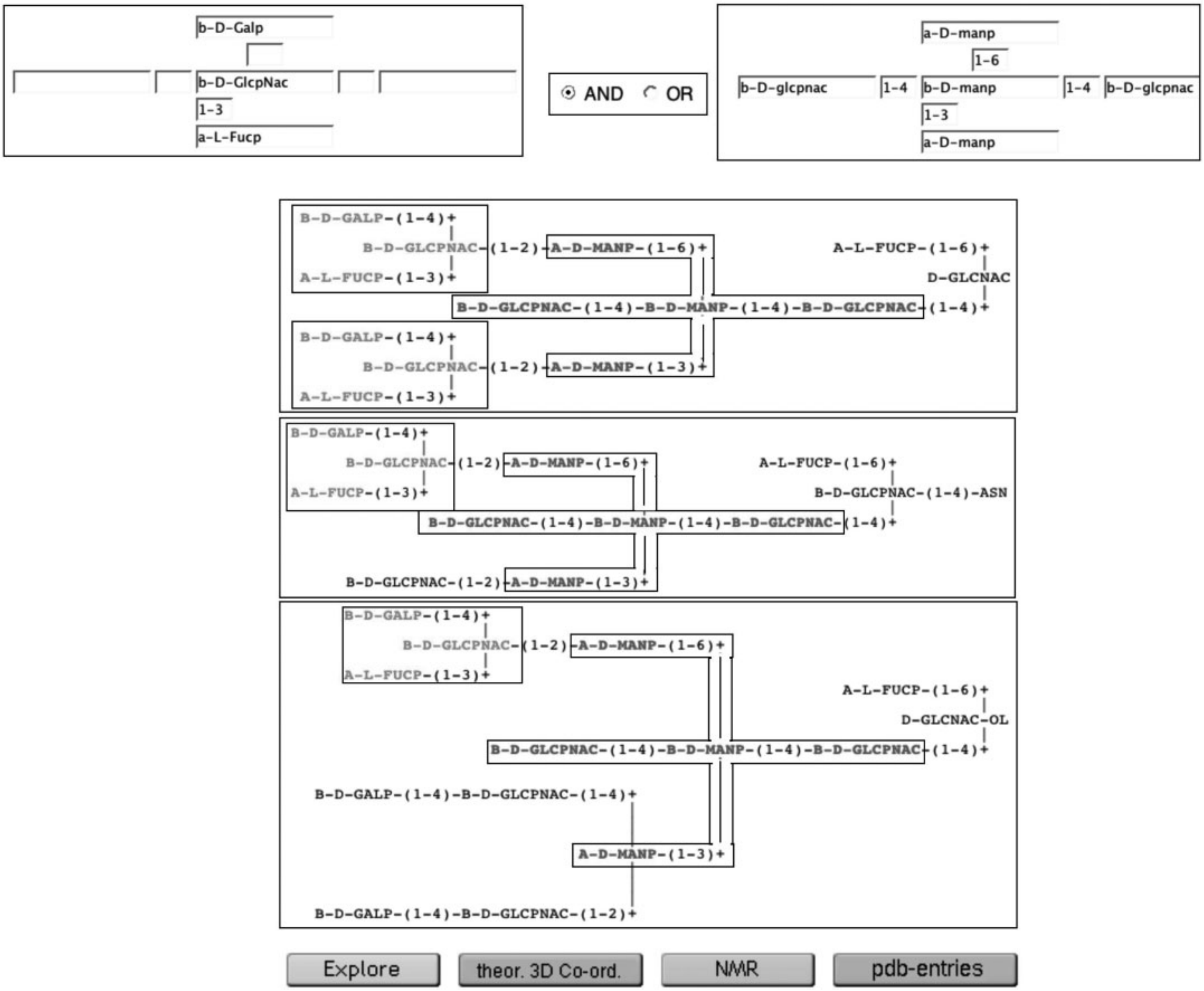

b-D-Galp
b-D-GlcpNac
1-3
a-L-Fucp
◉ AND ○ OR
a-D-manp
1-6
b-D-glcpnac
1-4
b-D-manp
1-4
b-D-glcpnac
1-3
a-D-manp
B-D-GALP-(1-4)+
B-D-GLCPNAC-(1-2)+A-D-MANP-(1-6)+
A-L-FUCP-(1-3)+
A-L-FUCP-(1-6)+
D-GLCNAC
B-D-GLCPNAC-(1-4)-B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)+
B-D-GALP-(1-4)+
B-D-GLCPNAC-(1-2)+A-D-MANP-(1-3)+
A-L-FUCP-(1-3)+
B-D-GALP-(1-4)+
B-D-GLCPNAC-(1-2)+A-D-MANP-(1-6)+
A-L-FUCP-(1-3)+
A-L-FUCP-(1-6)+
B-D-GLCPNAC-(1-4)-ASN
B-D-GLCPNAC-(1-4)-B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)+
B-D-GLCPNAC-(1-2)+A-D-MANP-(1-3)+
B-D-GALP-(1-4)+
B-D-GLCPNAC-(1-2)+A-D-MANP-(1-6)+
A-L-FUCP-(1-3)+
A-L-FUCP-(1-6)+
D-GLCNAC-OL
B-D-GLCPNAC-(1-4)-B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)+
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-4)+
A-D-MANP-(1-3)+
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)+
Explore
theor. 3D Co-ord.
NMR
pdb-entries

Fig. 2. Example for an advanced substructure search, where the retrieved entries contain both substructures given at the top. The structures found are indicated through different boxes. Also wildcards are supported.

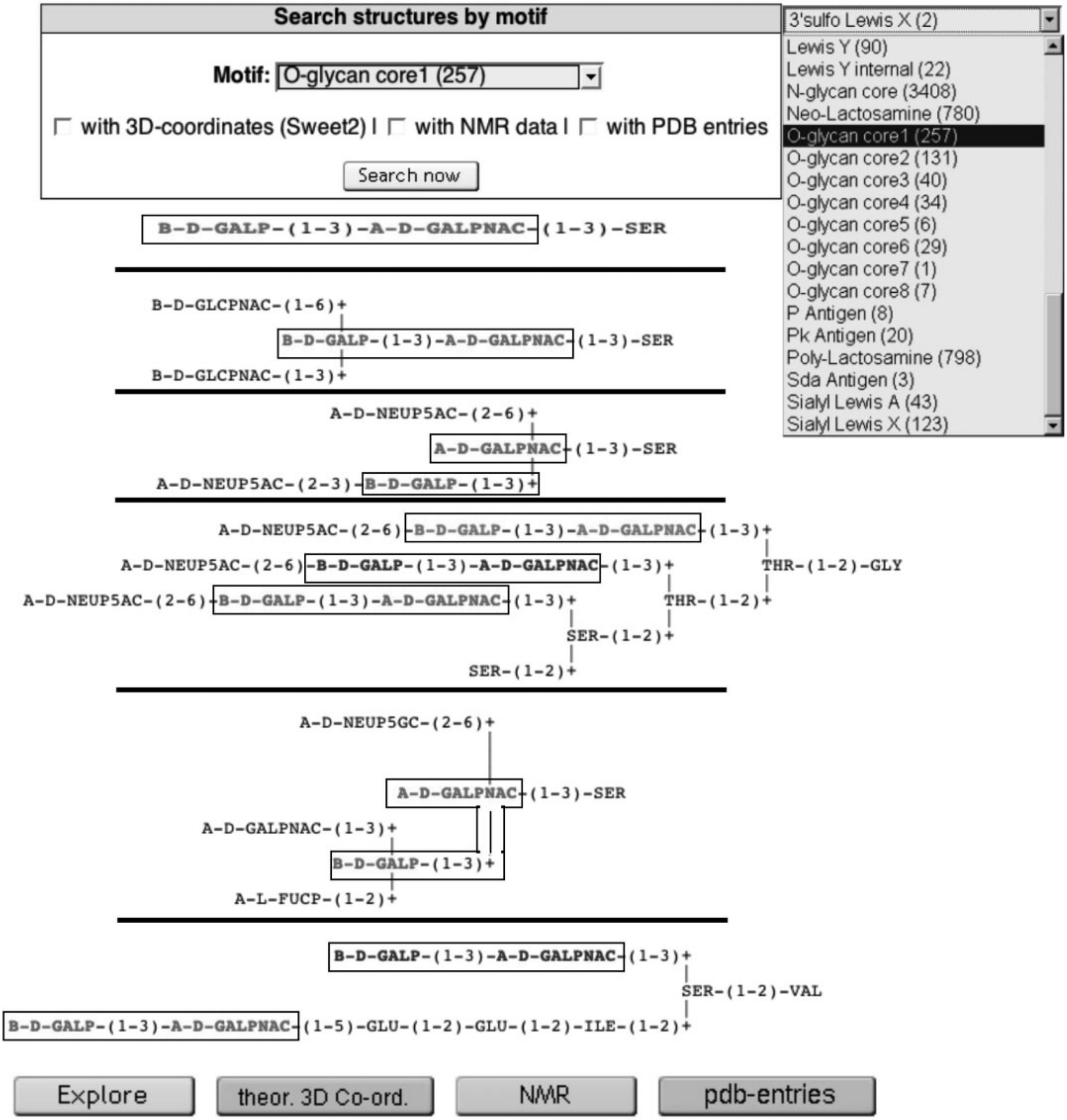
is applied to the input structure, all stored shifts values for a given code are recalled and a statistical analysis is performed. Normally, the mean of all stored shifts is taken for the prediction. The mean, minimum, maximum and standard deviation for the distribution of shifts is displayed for each estimated shift. The user has the possibility to examine the underlying distribution of shifts in detail and retrieve the original spectra from which specific shifts originate.

MS

To assist the interpretation of MS spectra, the total mass and all possible fragmentations of the glycosidic linkage are displayed when activating the corresponding button (Figure 5e). By default all B, C and X, Y fragments are presented. The other theoretically possible fragments can be easily looked up when posting the structure to the *GlycoFragment* (Lohmann and von der Lieth, 2003) service. Here, also A and Z fragments, various ions, different modifications like permethylation, peracetylation, and several anomeric attachments can be easily included in the calculation of fragments (see supplementary material 9 and examples listed when opening the *GlycoFragment* tool).

Experimental 3D structures

The PDB (Berman *et al.*, 2000) is the largest repository of experimentally determined 3D structures. About 5% of all entries contain also 3D coordinates for covalently attached *N*- or *O*-glycans or for noncovalently bound carbohydrates. The access to carbohydrate structures contained in PDB entries is accomplished through the *pdb2linucs* (Lütteke *et al.*, 2004) service (Figure 5f). It automatically detects



Search structures by motif

Motif: O-glycan core1 (257)

□ with 3D-coordinates (Sweet2) | □ with NMR data | □ with PDB entries

Fig. 3. Motif search: Structures comprising the *O*-glycan core 1 are shown.

ligands and glycan chains by analyzing 3D coordinates and their connectivity and converts the detected sugar structures to the LINUCS notation. This automatic encoding enables an easy integration of the detected glycans into the GLYCOSCIENCES.de portal. Because the PDB is weekly updated and the assignment of glycan structures can be performed in a semiautomatic way, which requires only modest supervision by a human expert, the user will always have access to all available glycan-related data in PDB. Figure 5f shows the first 10 of 17 PDB entries, which are currently available for Man_3. Activation of the explore button for a retrieved entry leads to a page with detailed information on the respective PDB entry (Figure 5g top). There, the *pdb2linucs* button displays the protein (in a cartoon

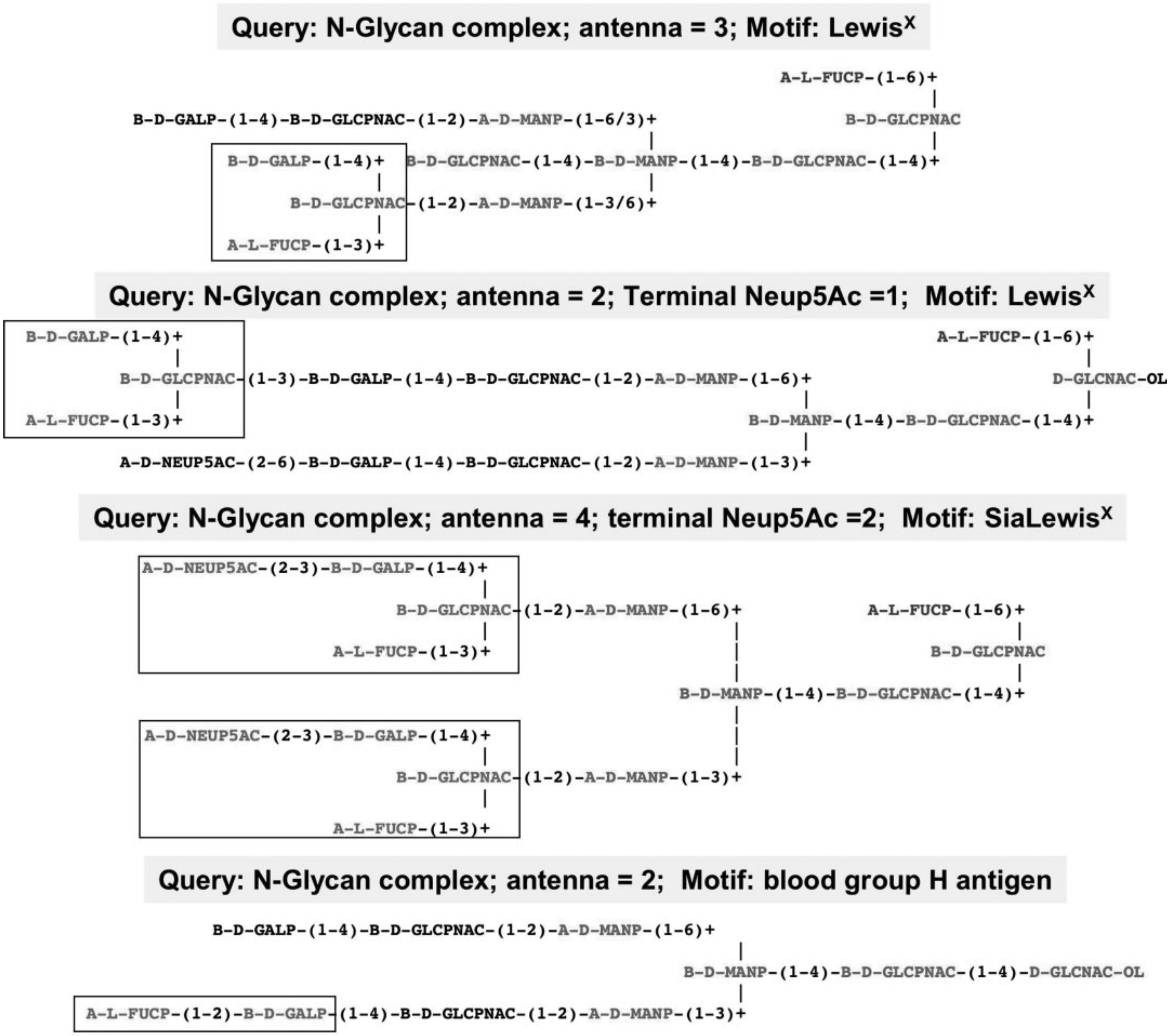
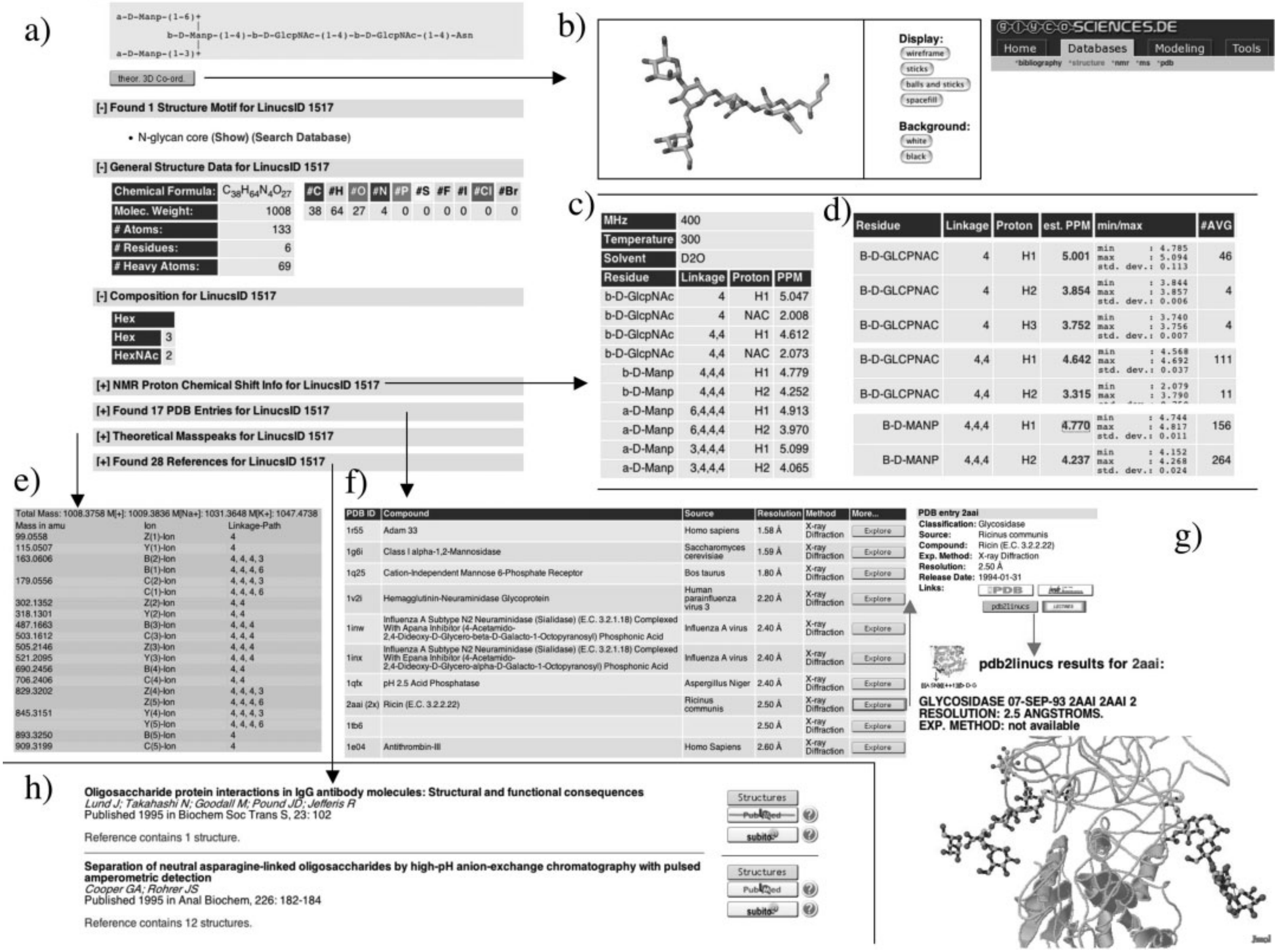

Query: N-Glycan complex; antenna = 3; Motif: Lewis^{X}
A-L-FUCP-(1-6)+
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)-A-D-MANP-(1-6/3)+
B-D-GLCPNAC
B-D-GALP-(1-4)+
B-D-GLCPNAC-(1-4)-B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)+
B-D-GLCPNAC-(1-2)-A-D-MANP-(1-3/6)+
A-L-FUCP-(1-3)+
Query: N-Glycan complex; antenna = 2; Terminal Neup5Ac =1; Motif: Lewis^{X}
B-D-GALP-(1-4)+
A-L-FUCP-(1-6)+
B-D-GLCPNAC-(1-3)-B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)-A-D-MANP-(1-6)+
D-GLCNAC-OL
A-L-FUCP-(1-3)+
B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)+
A-D-NEUP5AC-(2-6)-B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)-A-D-MANP-(1-3)+
Query: N-Glycan complex; antenna = 4; terminal Neup5Ac =2; Motif: SiaLewis^{X}
A-D-NEUP5AC-(2-3)-B-D-GALP-(1-4)+
B-D-GLCPNAC-(1-2)-A-D-MANP-(1-6)+
A-L-FUCP-(1-6)+
A-L-FUCP-(1-3)+
B-D-GLCPNAC
B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)+
A-D-NEUP5AC-(2-3)-B-D-GALP-(1-4)+
B-D-GLCPNAC-(1-2)-A-D-MANP-(1-3)+
A-L-FUCP-(1-3)+
Query: N-Glycan complex; antenna = 2; Motif: blood group H antigen
B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)-A-D-MANP-(1-6)+
B-D-MANP-(1-4)-B-D-GLCPNAC-(1-4)-D-GLCNAC-OL
A-L-FUCP-(1-2)-B-D-GALP-(1-4)-B-D-GLCPNAC-(1-2)-A-D-MANP-(1-3)+

Fig. 4. The *N*-glycan search option combined with the motif search. The search queries are given at the top. Only one typical representative structure is shown. The motifs found are indicated by boxes.

representation) with all attached *N*- and *O*-glycans as well as the noncovalently bound ligands using different color codes (Figure 5g bottom). Again, *JMOL* is used for the 3D display.

Several derived data sets are generated from the detected structures and their protein environment

1. *GlyVicinity*: (Lütteke *et al*., 2005) All amino acids in the spatial vicinity of each carbohydrate moiety and the interacting atoms are stored in a separate file and can be analyzed in various ways with the help of the *GlyVicinity* interface.
2. *Glyseq*: (Lütteke *et al*., 2005) A statistical analysis of the frequency of amino acids found in the neighborhood of the detected N- and O-glycosylation sites can be performed.
3. *GlyTorsion*: (Lütteke *et al*., 2005) The torsion angles of the glycosidic linkages between two sugar rings, which dominantly determine the 3D shape of carbohydrate structures, are automatically detected for attached glycans as well as ligands and are stored in a separate file. They can be recalled to build 3D structures of attached glycan, as it is done in the *GlyProt* (Bohne-Lang and von der Lieth, 2005) (Table I) service, to be compared with theoretically calculated conformational maps (*GlycoMapsDB*) or to judge if a carbohydrate 3D structure is reasonable (*Carp*).



```
a-D-Manp-(1-6)+
              |
        b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc-(1-4)-Asn
              |
a-D-Manp-(1-3)+
```

[-] Found 1 Structure Motif for LinucsID 1517

- N-glycan core (Show) (Search Database)

[-] General Structure Data for LinucsID 1517

Chemical Formula:	$C_{38}H_{64}N_4O_{27}$
Molec. Weight:	1008
# Atoms:	133
# Residues:	6
# Heavy Atoms:	69

#C	#H	#O	#N	#P	#S	#F	#I	#Cl	#Br
38	64	27	4	0	0	0	0	0	0

[-] Composition for LinucsID 1517

Hex	
Hex	3
HexNAc	2

[+] NMR Proton Chemical Shift Info for LinucsID 1517

[+] Found 17 PDB Entries for LinucsID 1517

[+] Theoretical Masspeaks for LinucsID 1517

[+] Found 28 References for LinucsID 1517

MHz	400		
Temperature	300		
Solvent	D2O		
Residue	Linkage	Proton	PPM
b-D-GlcpNAc	4	H1	5.047
b-D-GlcpNAc	4	NAC	2.008
b-D-GlcpNAc	4,4	H1	4.612
b-D-GlcpNAc	4,4	NAC	2.073
b-D-Manp	4,4,4	H1	4.779
b-D-Manp	4,4,4	H2	4.252
a-D-Manp	6,4,4,4	H1	4.913
a-D-Manp	6,4,4,4	H2	3.970
a-D-Manp	3,4,4,4	H1	5.099
a-D-Manp	3,4,4,4	H2	4.065

Residue	Linkage	Proton	est. PPM	min/max	#AVG
B-D-GLCPNAC	4	H1	5.001	min : 4.785 max : 5.094 std. dev.: 0.113	46
B-D-GLCPNAC	4	H2	3.854	min : 3.844 max : 3.857 std. dev.: 0.006	4
B-D-GLCPNAC	4	H3	3.752	min : 3.740 max : 3.756 std. dev.: 0.007	4
B-D-GLCPNAC	4,4	H1	4.642	min : 4.568 max : 4.692 std. dev.: 0.037	111
B-D-GLCPNAC	4,4	H2	3.315	min : 2.079 max : 3.790	11
B-D-MANP	4,4,4	H1	4.770	min : 4.744 max : 4.817 std. dev.: 0.011	156
B-D-MANP	4,4,4	H2	4.237	min : 4.152 max : 4.268 std. dev.: 0.024	264

Total Mass: 1008.3758 M[+]: 1009.3836 M[Na+]: 1031.3648 M[K+]: 1047.4738

Mass in amu	Ion	Linkage-Path
99.0558	Z(1)-Ion	4
115.0507	Y(1)-Ion	4
163.0606	B(2)-Ion	4, 4, 4, 3
	B(1)-Ion	4, 4, 4, 6
179.0556	C(2)-Ion	4, 4, 4, 3
	C(1)-Ion	4, 4, 4, 6
302.1352	Z(2)-Ion	4, 4
318.1301	Y(2)-Ion	4, 4
487.1663	B(3)-Ion	4, 4, 4
503.1612	C(3)-Ion	4, 4, 4
505.2146	Z(3)-Ion	4, 4, 4
521.2095	Y(3)-Ion	4, 4, 4
690.2456	B(4)-Ion	4, 4
706.2406	C(4)-Ion	4, 4
829.3202	Z(4)-Ion	4, 4, 4, 3
	Z(5)-Ion	4, 4, 4, 6
845.3151	Y(4)-Ion	4, 4, 4, 3
	Y(5)-Ion	4, 4, 4, 6
893.3250	B(5)-Ion	4
909.3199	C(5)-Ion	4

PDB ID	Compound	Source	Resolution	Method	More...
1r55	Adam 33	Homo sapiens	1.58 Å	X-ray Diffraction	Explore
1g6i	Class I alpha-1,2-Mannosidase	Saccharomyces cerevisiae	1.59 Å	X-ray Diffraction	Explore
1q25	Cation-Independent Mannose 6-Phosphate Receptor	Bos taurus	1.80 Å	X-ray Diffraction	Explore
1v2i	Hemagglutinin-Neuraminidase Glycoprotein	Human parainfluenza virus 3	2.20 Å	X-ray Diffraction	Explore
1inw	Influenza A Subtype N2 Neuraminidase (Sialidase) (E.C. 3.2.1.18) Complexed With Apana Inhibitor (4-Acetamido-2,4-Dideoxy-D-Glycero-beta-D-Galacto-1-Octopyranosyl) Phosphonic Acid	Influenza A virus	2.40 Å	X-ray Diffraction	Explore
1inx	Influenza A Subtype N2 Neuraminidase (Sialidase) (E.C. 3.2.1.18) Complexed With Epana Inhibitor (4-Acetamido-2,4-Dideoxy-D-Glycero-alpha-D-Galacto-1-Octopyranosyl) Phosphonic Acid	Influenza A virus	2.40 Å	X-ray Diffraction	Explore
1qfx	pH 2.5 Acid Phosphatase	Aspergillus Niger	2.40 Å	X-ray Diffraction	Explore
2aai (2x)	Ricin (E.C. 3.2.2.22)	Ricinus communis	2.50 Å	X-ray Diffraction	Explore
1tb6			2.50 Å	X-ray Diffraction	Explore
1e04	Antithrombin-III	Homo Sapiens	2.60 Å	X-ray Diffraction	Explore

PDB entry 2aai

Classification:	Glycosidase
Source:	Ricinus communis
Compound:	Ricin (E.C. 3.2.2.22)
Exp. Method:	X-ray Diffraction
Resolution:	2.50 Å
Release Date:	1994-01-31
Links:	

pdb2linucs results for 2aai:

GLYCOSIDASE 07-SEP-93 2AAI 2AAI 2
RESOLUTION: 2.5 ANGSTROMS.
EXP. METHOD: not available

Oligosaccharide protein interactions in IgG antibody molecules: Structural and functional consequences
Lund J; Takahashi N; Goodall M; Pound JD; Jefferis R
Published 1995 in Biochem Soc Trans S, 23: 102

Reference contains 1 structure.

Separation of neutral asparagine-linked oligosaccharides by high-pH anion-exchange chromatography with pulsed amperometric detection
Cooper GA; Rohrer JS
Published 1995 in Anal Biochem, 226: 182-184

Reference contains 12 structures.

Fig. 5. Display of all data available for the Man_3 glycan (LINUCS-ID 1517). To enable an easy navigation, a survey of the available data (**a**) is presented. By default, only the glycan structure in *CarbBank* representation is displayed. The user has the possibility to display the associated data by clicking on the corresponding line. **a** shows, in addition to the structure, the contained motif, the composition, and the general structure data. A click on the "theor. 3D co-ord." button invokes the display of a spatial structure (**b**), which has been automatically generated based on the *CarbBank* notation using the *SWEET-II* (Bohne *et al.*, 1999) service. **c** depicts the stored NMR shifts, which are experimentally determined. **d** shows some shifts of Man_3 estimated with the NMR-prediction option. **e** displays the theoretically calculated fragments resulting from cleavages of the glycosidic bonds (B-, C- and Z-, Y ions). **f** represents a list of PDB entry codes, the protein name, its biological source as well as resolution, which contain Man_3. A click on the corresponding explore button (here the one for PDB entry 2aai) will recall supplementary data for a given PDB entry (**g top**). The *pdb2linucs* (Lütteke *et al.*, 2004) approach has been used to automatically identify the carbohydrates. The "pdb2linucs" button leads to a page (**g bottom**) displaying the coordinates using JMOL. **h** illustrates that also associated references, where Man_3 has been explicitly mentioned, can be retrieved. Here only two of 28 stored entries are displayed.

Last but not least, the *GLYCOSCENCES.de* portal provides access to all stored references for a given structure (Figure 5h). An automatic link to abstracts as provided by the *PubMed* service is enabled. However, since the systematic update of *CarbBank* stopped in 1998 and only references containing NMR peak-lists have been added, not the complete literature is covered.

Summary and discussion

Table I presents a list of the purposes of all applications accessible through the GLYCOSCIENCES.de portal. The structure-oriented integration of glycan-related scientific data originating from various resources can be regarded as a proof of principle, that the users' dream recently expressed at the joint meeting of the US and Japanese consortia for glycomics might become reality within a foreseeable future: "It was agreed that the construction of databases is an area of interest to everyone and is ripe for synergistic effort. A distributed database with a single search engine, accessible to everyone, was thought to be a worthy goal." (*CFG Consortium Quarterly Newsletter*, Vol. 3, No. 2, December 2004, http://www.functionalglycomics.org/static/consortium/news.shtml#newsletter).

Basic prerequisite for this idea is that the organizations maintaining digital collections of glycan-related scientific data make it universally and freely accessible through the

Table I. List of available glyco-related applications through the GLYCOSCIENCES.de portal.

Name	Purpose	Input	Concept	Reference
PDB-related services				
pdb2linucs	Extraction of carbohydrate structures from PDB files	PDB-ID or coordinates	Automatic assignment of ligand and glycan chains based on 3D coordinates	Lütteke *et al.* (2004)
GlyVicinity	Analysis of protein–carbohydrate interactions	Carbohydrate residue	Statistics of amino acids in spatial vicinity of carbohydrates	Lütteke *et al.* (2005)
GlySeq	Analysis of protein sequences around glycosylation sites	Type of glycosylation site (Asn, Ser, Thr,...)	Statistics of amino acids in the neighborhood of glycosylation sites	Lütteke *et al.* (2005)
GlyTorsion	Analysis of carbohydrate torsion angles	Disaccharide	Display of all available torsion angles for one glycosidic linkage, ring torsions, omega torsions, etc.	Lütteke *et al.* (2005)
Carp	Analysis of torsion angles of glycosidic linkages	PDB-ID or coordinates	Comparison of all available torsion angles with the ones found in a given PDB file	Lütteke *et al.* (2005)
pdb-care	Check of carbohydrate residue nomenclature	PDB-ID or coordinates	Comparison of names given in the PDB file with those generated with pdb2linucs	Lütteke and von der Lieth (2004)
Modelling tools				
SWEET-II	Quick generation of a reliable 3D conformation	Carbohydrate nomenclature	Linking of ready-made 3D molecular templates of monosaccharides. Subsequent optimizing of the 3D structure using the MM3 force field	Bohne *et al.* (1999)
Dynamic molecules	Exploring conformational space	Glycan	Internet portal which provides molecular dynamics simulations for oligosaccharides	Frank *et al.* (2003)
GlyProt	*In silico* glycosylation of proteins	PDB-ID plus glycan	3D structure of protein is required. Potential N-glycosylations site are automatically detected. To be attached glycans are constructed with SWEET-II	Bohne-Lang and von der Lieth (2005)
Spectroscopic tools				
Glycofragment	Calculates all theoretically possible fragments	Glycan	Finds the main fragments of glycans which occur in MS spectra	Lohmann and von der Lieth (2003)
GlycoProfiling	Finds all glycans with a given molpeak	Molpeak, ion and derivatization	Compares the molpeak with all m/z values contained in a database of *N*-glycans	Lohmann and von der Lieth (2004)
GlycoSearchMS	Finds glycans whose fragmentation pattern match best with the spectrum	List of *m*/*z* values, ion and derivatization	Compares the peak list with a list of theoretically calculated fragments derived from a database of N-glycan structures	Lohmann and von der Lieth (2004)
NMR spectrum search	Finds glycans whose spectrum matches best with the input spectrum	List of NMR-shifts (1H- or 13C)	Compares a list of NMR-shifts with all spectra contained in the database. Displays a hit list of spectra and structures in descending order of their spectral similarity	
NMR-spectrum estimation	Estimation of ^{1}H- or ^{13}C spectra : assumption : similar structural environments exhibit similar spectra	Glycan	NMR shift estimation is based on an appropriate encoding of the structural environment for each atom, which is stored together with the assigned shift value.	
Other tools				
LINUCS	Linear Notation for Unique description of Carbohydrate Structures	Glycan	Normalization of sugar topologies starting from the reducing end and using the linkage path for sorting	Bohne-Lang *et al.* (2001)
LiGraph	Schematic drawings of oligosaccharides are often used to display glycan structures	Glycan	Normalization of sugar topologies starting from the reducing end and using a set of topology based rules	
PubFinder	Search for thematically related references	Set of references	Automatic identification of PubMed abstracts that deal with a specific scientific subject. The search is based on a set of representative abstracts, which delineate well a certain scientific topic	Goetz and von der Lieth (2005)
GlycoMapsDB	Comparison of ready-made conformational maps with experimental data	Disaccharide	Conformational maps are automatically generated from long-term molecular dynamics simulation using Dynamic Molecules	

Internet (open access philosophy: *free availability and unrestricted use*) in an easily readable and agreed format without any barrier. Unfortunately this situation has not yet been reached and it will probably take some time to convince database providers to join the open access community and to invest time to convert their internal descriptions to standard formats.

It is obvious that the upcoming high-throughput glycoproteomics and glycomics projects will produce a large amount of experimental data. It is predictable that especially the availability of glyco-arrays providing data which oligosaccharide binds to which lectin as well as the MS-based profiling of the glycans, which are found in normal and diseased tissues, will attract considerable attention. Because, for example, the CFG has clearly announced that they will provide open access to their data and since most of the information is already available in a digital format on the Internet, it will be a logic expansion of the GLYCOSCIENCES.de portal to cross-link this information with the already available data. A prerequisite for an efficient exchange of data will be the agreement to a generally accepted exchange format as well as to a common software interface. The need to establish a readable exchange format has been recognized by various research groups: CabosML (Kikuchi *et al.*, 2005), GLYcan Data Exchange (GLYDE) (http://lsdis.cs.uga.edu/projects/glycomics/index.php?page=4) and Bacterial Carbohydrate Structure Data Base (http://www.glyco.ac.ru/bcsdb/start.shtml).

Consequently, several proposals for an eXtensible Markup Language (XML)-based description of glycan structures already exist and it is foreseeable that some form of consensus will be reached within the near future. This agreement may not cover all peculiarities for all types of carbohydrates found in nature. However, it will be sufficiently comprehensive for all glycan structures found in mammalians. It seems that the simple object access protocol (SOAP) (http://www.w3.org/TR/soap/) is now the broadly accepted procedure for communication between applications. Being designed to communicate through Internet, it is well suited to be also used for the exchange of glycan-related data between distributed computers. Taking together, the time seems to be mature to establish an online connection of distributed databases at least between the larger already established projects.

Besides the larger projects, many other initiatives exist where glycoscientists have made available their scientific data on the Internet, using a representation that is often closely related to a particular experimental data format or to a certain biological system (von der Lieth, 2004). This diversity of data models hampers an efficient cross-linking with other resources and makes comparative glycomics analysis difficult. Currently, the creation of disconnected and incompatible islands of glycomics data continues.

One obvious reason for this undesirable situation is that no well-established procedures and repositories exist, where glycoscientists can deposit their glycan structures and related experimental data (von der Lieth, 2004). In genomics and proteomics research, it is a standard procedure that scientists submit their sequences before publication to one of the large repositories like GenBank (Benson *et al.*, 2004) for DNA data or PDB (Berman *et al.*, 2004) for 3D macromolecular structures. No such procedures exist for glycomics data. Although the currently produced amount of scientific data as compared with genomics and proteomics is still modest in size, the foreseeable rapid evolution of glycomics research will result in an increasing number of glycan structures found in various tissues and species. Therefore, it is an obvious demand that a corresponding repository will have to be established for glycomics research as well. However, since glycans exhibit branched structures and the monosaccharide units are connected in several ways, oligosaccharides cannot be encoded using a simple linear code. Therefore, additional software tools have to be provided, which enable all scientists to input glycan structures in an easily manageable way.

Based on the idea that the Internet offers the unique chance to constitute a global and interactive communication for scientific data, the EUROCarbDB (http://www.eurocarbdb.org) project, a design study funded by the sixth framework program of the European Union, develops a network of locally installed distributed databases for glycosciences. The availability of such tools will encourage people to input their recorded experimental data into a local database that may be kept private until it is published. Additionally, the released data will be further annotated, stored and archived in a central database which will be maintained at the European Bioinformatics Institute (EMBL-EBI) (http://www.ebi.ac.uk).

The existence of a broadly accepted open access repository for glyco-related experimental data will hopefully secure that the loss of primary data in glycobiology research will be considerably reduced. The agreement to quality standards and notations will not only raise the scientific usefulness of the stored data, it will open the opportunity to apply various data-mining approaches including multivariate statistics and artificial neural network algorithms to extract new information and to derive new knowledge which until now is hidden in unstructured data.

Supplementary data

Supplementary data are available at *Glycobiology* online (http://glycob.oxfordjournals.org/).

Acknowledgments

The development of the GLYCOSCIENCES.de portal is funded by a grant from the German Research Council (Deutsche Forschungsgemeinschaft: DFG) within the digital library program.

Abbreviations

3D, three-dimensional; CCSD, Complex Carbohydrate Structure Database; CFG, Consortium for Functional Glycomics; KEGG, Kyoto Encyclopedia of Genes and Genomes; LINUCS, LInear Notation for Unique description of Carbohydrate Sequences; MS, mass spectrometry; NMR, nuclear magnetic resonance; PDB, Protein Database.

References

Aoki, K., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M., and Mamitsuka, H. (2003) Efficient tree-matching methods for accurate carbohydrate database queries. *Genome Inform.*, **14**, 134–143.

Aoki, K., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S., and Kanehisa, M. (2004) KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res.*, **32**, W267–W272.

Aoki, K., Mamitsuka, H., Akutsu, T., and Kanehisa, M. (2005) A score matrix to reveal the hidden links in glycans. *Bioinformatics*, **21**, 1457–1463.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Blixt, O., Head, S., Mondala, T., Scanlan, C., Huflejt, M., Alvarez, R., Bryan, M., Fazio, F., Calarese, D., Stevens, J., and others. (2004) Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 17033–17038.

Bohne, A., Lang, E., and von der Lieth, C.W. (1999) SWEET – WWW-based rapid 3D construction of oligo- and polysaccharides. *Bioinformatics*, **15**, 767–768.

Bohne-Lang, A. and von der Lieth, C.W. (2005) GlyProt: *is silico* glycosylation of proteins. *Nucleic Acids Res.*, **33**, W214–W219.

Bohne-Lang, A., Lang, E., Forster, T., and von der Lieth, C.W. (2001) LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr. Res.*, **336**, 1–11.

Cooper, C., Joshi, H., Harrison, M., Wilkins, M., and Packer, N. (2003) GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.*, **31**, 511–513.

Dell, A. and Morris, H. (2001) Glycoprotein structure determination by mass spectrometry. *Science*, **291**, 2351–2356.

Doubet, S. and Albersheim P. (1992) CarbBank. *Glycobiology*, **2**, 505.

Doubet, S., Bock, K., Smith, D., Darvill, A., and Albersheim, P. (1989) The complex carbohydrate structure database. *Trends Biochem. Sci.*, **14**, 475–477.

Feizi, T. and Chai, W. (2004) Oligosaccharide microarrays to decipher the glyco-code. *Nat. Rev. Mol. Cell Biol.*, **5**, 582–588.

Frank, M., Gutbrod, P., Hassayoun, C., and von der Lieth, C. (2003) Dynamic molecules: molecular dynamics for everyone. An internet-based access to molecular dynamic simulations: basic concepts. *J. Mol. Model*, **9**, 308–315.

Goetz, T. and von der Lieth, C. (2005) PubFinder: a tool for improving retrieval rate of relevant PubMed. *Nucleic Acids Res.*, **33**, W774–W778.

Goetz, T., Bohne-Lang, A., Frank, M., Lohmann, K., Loss, A., Lütteke, T., and von der Lieth, C.W. (2004) Glycosciences.de: an Internet portal for glyco-related data from open access resources. In Giegerich R., Stoye, J. (eds.), *German Conference on Bioinformatics*. Köllen Druck+ Verlag GmbH, Bonn, Germany, pp. 115–119.

Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K., Ueda, N., Hamajima, M., Kawasaki, T., and Kanehisa, M. (2005) KEGG as a glycome informatics resource. *Glycobiology*. Epub ahead of print.

Kikuchi, N., Kameyama, A., Nakaya, S., Ito, H., Sato, T., Shikanai, T., Takahashi, Y., and Narimatsu, H. (2005) The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics*, **21**, 1717–1718.

Lohmann, K. and von der Lieth, C. (2003) GLYCO-FRAGMENT. A web tool to support the interpretation of mass spectra of complex carbohydrates. *Proteomics*, **3**, 2028–2035.

Lohmann, K. and von der Lieth, C. (2004) GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res.*, **32**, W261–W266.

Loss, A., Bunsmann, P., Bohne, A., Loss, A., Schwarzer, E., Lang, E., and von der Lieth, C.W. (2002) SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res.*, **30**, 405–408.

Lütteke, T. and von der Lieth, C.W. (2004) pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics*, **5**, 69.

Lütteke, T., Frank, M., and von der Lieth, C.W. (2004) Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr. Res.*, **339**, 1015–1020.

Lütteke, T., Frank, M., and von der Lieth, C.W. (2005) Carbohydrate structure suite (CSS): analysis of carbohydrate, 3D structures derived from the PDB. *Nucleic Acids Res.*, **33**, D242–D246.

Pratt, M.R. and Bertozzi, C.R. (2005) Synthetic glycopeptides and glycoproteins as tools for biology. *Chem. Soc. Rev.*, **34**, 58–68.

Sayle, R. and Milner-White, E. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.

Sequencing, I.H.G. (2004) Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

van Kuik, J. and Vliegenthart, J.(1992) Databases of complex carbohydrates. *Trends Biotechnol.*, **10**, 182–185.

van Kuik, J., Hard, K., and Vliegenthart, J.A. (1992) 1H NMR database computer program for the analysis of the primary structure of complex carbohydrates. *Carbohydr. Res.*, **235**, 53–68.

von der Lieth, C.W. (2004) An endorsement to create open databases for analytical data of complex carbohydrates. *J. Carbohydr. Chem.*, **23**, 277–297.

von der Lieth, C.W., Bohne-Lang, A., Lohmann, K., and Frank, M. (2004) Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform.*, **5**, 164–178.

Wong, C. (2005) Protein glycosylation: new challenges and opportunities. *J. Org. Chem.*, **70**, 4219–4225.