Published online 2 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D373–D376***
doi:10.1093/nar/gkq1014

GlycomeDB—a unified database for carbohydrate structures

René Ranzinger*, Stephan Herget, Claus-Wilhelm von der Lieth and Martin Frank

German Cancer Research Center (DKFZ), Core Facility: Molecular Structure Analysis (W160), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

Received August 15, 2010; Revised October 4, 2010; Accepted October 8, 2010

ABSTRACT

GlycomeDB integrates the structural and taxonomic data of all major public carbohydrate databases, as well as carbohydrates contained in the Protein Data Bank, which renders the database currently the most comprehensive and unified resource for carbohydrate structures worldwide. GlycomeDB retains the links to the original databases and is updated at weekly intervals with the newest structures available from the source databases. The complete database can be downloaded freely or accessed through a Web-interface (www.glycome-db.org) that provides flexible and powerful search functionalities.

INTRODUCTION

In a recent NIH whitepaper (1) the lack of a comprehensive, curated carbohydrate structure database was identified as the largest deficit in glycomics and glycobiology research. The Complex Carbohydrate Structure Database (CCSD) (2), initiated in the 1980s, was the largest effort to date to collect carbohydrate structures, mainly through retrospective manual extraction from the literature. The database contained about 50 000 entries when it ceased to be updated in the late 1990s due to a lack of funding. Since then different specialized databases have been developed, which were initially seeded with a subset of the structures contained in the CCSD (3). Subsequently these databases were further extended with carbohydrate structures reflecting the research focus of the group that maintained the database. As a result, different valuable collections of carbohydrate data have emerged over recent years, for example: the Bacterial Carbohydrate Structure Database (BCSDB) (4) that collects all published bacterial carbohydrate structures (including their NMR spectra); the database of the Consortium for Functional Glycomics (CFG) that provides access to primary experimental data like that from glycan microarray screens (5); and the Kyoto Encyclopedia of Genes and Genomes (KEGG) that contains glycan-related biosynthetic pathways (6). Unfortunately each of these databases uses a different 'sequence format' for encoding carbohydrate structures, making it difficult to query across all public databases and analyze or compare their content, or simply to find out whether some additional information on a particular carbohydrate structure is available in any of the databases.

GlycomeDB—SCOPE AND IMPLEMENTATION

In 2005, a new initiative was begun to overcome the isolation of the public carbohydrate structure databases and to create a comprehensive index of all available structures with cross-links back to the original databases. To achieve this goal, structures of the freely available databases were translated to the GlycoCT sequence format (7), if possible, and stored in a new database, the GlycomeDB (8). The integration process is performed incrementally on a weekly basis, updating the GlycomeDB with the newest structures available in the associated databases. A JAVA software application called GlycoUpdateDB, which is complemented by a PostgreSQL database, is used to download the data from the public databases, reads their sequence notations and translates them to the GlycoCT encoding format. In addition, the taxonomic annotations are standardized semi-automatically based on curated tables that map the (free-text) annotations used in the source databases to NCBI taxonomy IDs [for more details see (8)]. To extract the carbohydrate structures from the Protein Data Bank (PDB) the *pdb2linux* tool is used (9). During the integration process automated checks are performed; structures that contain errors are reported to the administrators of the original database.

*To whom correspondence should be addressed. Tel: +706 542 4401; Fax: +706 542 4412; Email: rene.ranzinger@glycome-db.org
Present address:
René Ranzinger, Complex Carbohydrate Research Center, The University of Georgia, 315 Riverbend Road, Athens, GA 30602, USA.

A major challenge during the initial integration process was the lack of a controlled vocabulary for carbohydrate and non-carbohydrate residue names. Even within a single database the same monosaccharide could have different names. In total 12 253 different residues names were extracted from the sequences stored in the original carbohydrate databases, 5854 of which were identified as non-carbohydrate residues, mainly aglycons, such as amino acids, lipids or other small organic molecules attached to the reducing end of the carbohydrate. In total 5330 residue names could be identified as monosaccharides and were assigned a standardized GlycoCT encoding. The remaining 1069 residue names could not be interpreted so far. Based on the initial analysis of the namespace used to encode carbohydrate structures in the various databases, a dictionary has been created that contains mappings of the various encoding formats. The dictionary is now used to support the automated update process. If a new residue name appears, this is reported to the database curator who can then check whether the residue name is valid and include the new residue into the dictionary. Finally, a web interface has been developed (www.glycome-db.org) as a single query point for all open access carbohydrate structure databases (10).

DATABASE CONTENT

GlycomeDB contains the unified carbohydrate sequences of all publicly accessible databases that contain carbohydrates structures. In total 121 766 original sequences were parsed and integrated. Currently (August 2010) there are 35 873 unique carbohydrate sequences—with taxonomic annotations if available—stored in GlycomeDB, 11 822 of which are fully determined carbohydrates. A carbohydrate structure is defined as 'fully determined' if all monosaccharide characteristics (base type, anomer, ring size, substituents, modifications, etc) and all linkage positions are known. For polysaccharides the number of repeating units needs to be determined as well. An overview of the number of carbohydrate structures contributed by each database is given in Table 1.

Data retrieval and presentation

Four major structural query options are implemented in GlycomeDB, namely 'exact structure search', 'substructure search', 'similarity search' and 'maximum common substructure search' (10). Structural queries can be entered graphically, either using GlycanBuilder (14) as the default, or using DrawRINGS, developed by a Japanese group at SOKA University, Tokyo (http://rings.t.soka.ac.jp). It is also possible to specify the query structure by using different machine-readable encoding formats, among which are CarbBank format (2), LINUCS (15), LinearCode® (16), BCSDB encoding (4) and Glyde II (http://glycomics.ccrc.uga.edu/core4/informatics-glyde-ii.html).

Next to the exact structure search, which is based on a comparison of ordered GlycoCT encodings (7), it is possible to generate queries with partially unknown information on the monosaccharide level, i.e. unknown anomeric center, ring size, or absolute configuration. It is also possible to restrict the search to specific taxonomic sources, as GlycomeDB applies consistently the NCBI taxonomy for the taxonomic data (17). The various search options can be combined sequentially to a multistep query refinement workflow, which allows very complex queries to be performed.

Using the GlycomeDB information page for individual structures (Figure 1), the user can use hyperlinks to navigate to the relevant pages of the external databases, which offer additional information such as literature references, experimental data or 3D structures. Additionally, information about bound aglycons and structural motifs, and a selectable sequence encoding are displayed. For more detailed information about the various aglycons attached to a particular carbohydrate, the user is guided to the original databases by following the link 'Show remote structure evidences'.

SUMMARY AND OUTLOOK

GlycomeDB integrates the structural and taxonomic data of all major public carbohydrate databases, as well as carbohydrates contained in the Protein Data Bank, which

Table 1. Overview of the number of original unique carbohydrate or glycoconjugate sequences contained in the source databases (encoded in the database-specific format, including the aglycon unit) and the number of unique GlycoCT sequences generated after removing the aglycon and parsing the remaining code

External database	Number of sequences in external database	Number of unique GlycoCT sequences	Fully determined carbohydrate sequences	URL
BCSDB (4)	8119	6536 (4149)	1972 (1277)	http://www.glyco.ac.ru/bcsdb3/
CCSD (2)	23 402	14 887 (1544)	7406 (462)	http://www.genome.jp/dbget-bin/www_bfind?carbbank
CFG (5)	8873	6285 (4143)	397 (110)	http://www.functionalglycomics.org/
EUROCarbDB	13 467	13 308 (411)	8924 (139)	http://www.ebi.ac.uk/eurocarb/
Glycobase(Lille) (11)	247	197 (145)	195 (143)	http://glycobase.univ-lille1.fr/base/
GLYCOSCIENCES.de (12)	23 285	15 829 (391)	9225 (36)	http://www.glycosciences.de/
KEGG (6)	10 969	10 160 (6128)	1610 (179)	http://www.genome.jp/kegg/glycan/
PDB (13)	905	733 (0)	708 (0)	http://www.rcsb.org/pdb/

The numbers in brackets denote the number of sequences that are stored exclusively in this database. Currently GlycomeDB contains 35 873 unique carbohydrate sequences and 11 822 fully determined carbohydrate sequences. See text for the criteria of a 'fully determined sequence'.

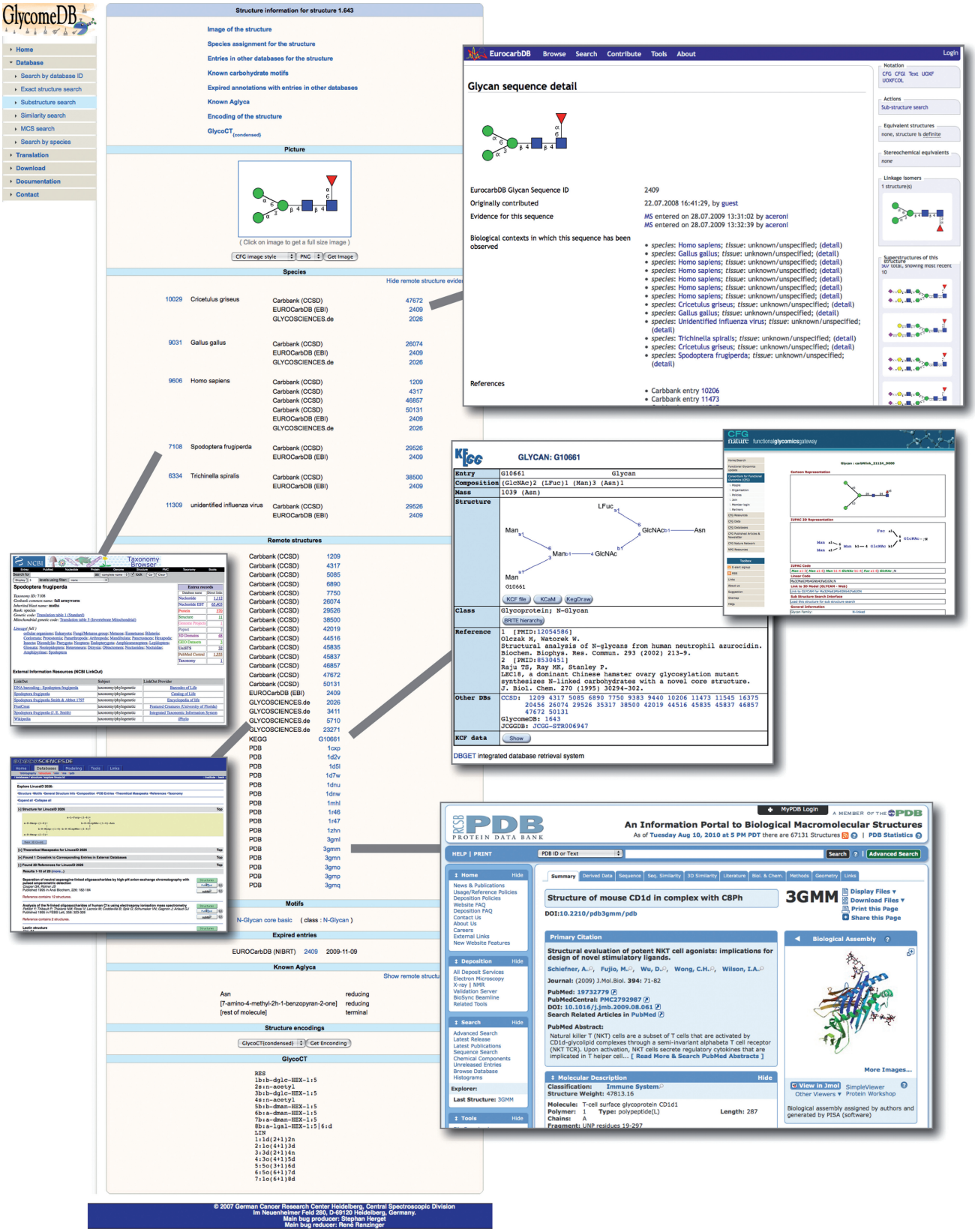

Figure 1. The structure information page of GlycomeDB. The carbohydrate is displayed in CFG-style cartoon representation. Species annotations and hyperlinks to external databases are available.

renders the database currently the most comprehensive and unified resource for carbohydrate structures worldwide. Hyperlinks to the original source of the data are established, so users can use the GlycomeDB Web-portal to access efficiently relevant additional information, which is only available in the original databases. GlycomeDB is a database that integrates knowledge from other existing databases, therefore only carbohydrate structures that are stored in any of these databases will be integrated and cross-linked in GlycomeDB. Unfortunately, GlycomeDB cannot provide access to all published structures because, in contrast to proteomics and genomics, in glycomics there is not yet a procedure established that requires deposition of new structures in the context of publication. Therefore it can be assumed that not all published structures are currently available in a database. However, if a public database will be used in the future to deposit systematically new structures, these structures should also be automatically available in GlycomeDB. In general, the quality of the data depends on the quality of the referenced databases and their curation processes. Nevertheless GlycoUpdateDB applies additional validation checks during the integration process in order to improve the quality of the data. The curated database can be downloaded and used freely by interested scientists. It can be assumed that the development of annotation tools in MS and NMR that require a library of existing carbohydrate structures as reference data will benefit from the availability of GlycomeDB. Additionally, the data contained in GlycomeDB can facilitate statistical analyses of the 'glycospace' of different organisms (18,19).

AVAILABILITY

GlycomeDB can be accessed using a Web-portal (http://www.glycome-db.org/) or the complete database can be downloaded as a compressed zip archive, containing all structures that have been integrated (http://www.glycome-db.org/downloads/). The structures are stored in regular XML files according to the Glyde II specification and can be used by any software that supports this format.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to the EU (6th Research Framework Program, RIDS). The authors wish to thank the maintainers of all public glycomics database projects for their cooperation and helpfulness. This project would be unthinkable without their support. We also would like to thank all our collaborators from the EUROCarbDB project.

FUNDING

EU (6th Research Framework Program, RIDS contract number 011952); German Research Foundation (DFG BIB 46 HDdkz 01-01). Funding for open access charge: German Cancer Research Center (DKFZ), Heidelberg, Germany.

Conflict of interest statement. None declared.

REFERENCES

1. Packer,N.H., von der Lieth,C.-W., Aoki-Kinoshita,K.F., Lebrilla,C.B., Paulson,J.C., Raman,R., Rudd,P., Sasisekharan,R., Taniguchi,N. and York,W.S. (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 (11–13 September 2006). *Proteomics*, **8**, 8–20.
2. Doubet,S., Bock,K., Smith,D., Darvill,A. and Albersheim,P. (1989) The complex carbohydrate structure database. *Trends Biochem. Sci.*, **14**, 475–477.
3. Ranzinger,R., Herget,S., Lutteke,T. and Frank,M. (2009) Carbohydrate structure databases. In Cummings,R.D. and Pierce,J.M. (eds), *Handbook of Glycomics*. Elsevier, Amsterdam, pp. 211–233.
4. Toukach,F.V. (2009) Bacterial carbohydrate structure database version 3. *Glycoconjugate J.*, **26**, 856.
5. Raman,R., Venkataraman,M., Ramakrishnan,S., Lang,W., Raguram,S. and Sasisekharan,R. (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, **16**, 82R–90R.
6. Hashimoto,K., Goto,S., Kawano,S., Aoki-Kinoshita,K.F., Ueda,N., Hamajima,M., Kawasaki,T. and Kanehisa,M. (2006) KEGG as a glycome informatics resource. *Glycobiology*, **16**, 63R–70R.
7. Herget,S., Ranzinger,R., Maass,K. and von der Lieth,C.-W. (2008) GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr. Res.*, **343**, 2162–2171.
8. Ranzinger,R., Herget,S., Wetter,T. and von der Lieth,C.-W. (2008) GlycomeDB - integration of open-access carbohydrate structure databases. *BMC Bioinformatics*, **9**, 384.
9. Lutteke,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.
10. Ranzinger,R., Frank,M., von der Lieth,C.W. and Herget,S. (2009) Glycome-DB.org: a portal for querying across the digital world of carbohydrate sequences. *Glycobiology*, **19**, 1563–1567.
11. Maes,E., Bonachera,F., Strecker,G. and Guerardel,Y. (2009) SOACS index: an easy NMR-based query for glycan retrieval. *Carbohydr. Res.*, **344**, 322–330.
12. Lutteke,T., Bohne-Lang,A., Loss,A., Goetz,T., Frank,M. and von der Lieth,C.-W. (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology*, **16**, 71R–81R.
13. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
14. Ceroni,A., Dell,A. and Haslam,S.M. (2007) The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol. Med.*, **2**, 3.
15. 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.
16. Banin,E., Neuberger,Y., Altshuler,Y., Halevi,A., Inbar,O., Nir,D. and Dukler,A. (2002) A novel Linear Code((R)) nomenclature for complex carbohydrates. *TIGG*, **14**, 127–137.
17. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
18. Werz,D.B., Ranzinger,R., Herget,S., Adibekian,A., von der Lieth,C.-W. and Seeberger,P.H. (2007) Exploring the structural diversity of mammalian carbohydrates ("glycospace") by statistical databank analysis. *ACS Chem. Biol.*, **2**, 685–691.
19. Herget,S., Toukach,P.V., Ranzinger,R., Hull,W.E., Knirel,Y.A. and 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.