glycosciences.de: An Internet Portal for Glyco-related Data from Open Access Resources*

T. Götz, A. Bohne-Lang, M. Frank, K. Lohmann, A. Loss, T. Lütteke, C.-W. von der Lieth

German Cancer Research Center (DKFZ), Central Spectroscopic Department B090
Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany
email: t.goetz@dkfz.de, a.bohne@dkfz.de, w.vonderlieth@dkfz.de

Abstract: Carbohydrates are involved in a variety of fundamental biological processes, e.g. cellular differentiation, embryonic development or fertilization. Since glycans are secondary gene products and their structure cannot be easily predicted from DNA sequences, very little of the bioinformatics algorithm and techniques developed for genomics and proteomics research can be directly adapted for glycomics. The development of new and advanced bioinformatics tools, algorithms and data collections for glycobiology is an absolute requirement to manage and analyze successfully the large amount of data which will be produced by the upcoming glycomics project.

The *glycosciences.de* portal aims to provide a platform, where (a) various glyco-related data collections originating from diverse open access resources as well as (b) tools helping to interpret and analyse experimental data of complex carbohydrates are made available using standard internet technologies. Access is provided based either on (sub)-structural descriptions, bibliographic data, physicochemical properties and experimental data like NMR, chemical shifts and MS-spectra. A normalised description for carbohydrate structures called LINUCS (**LI**near **N**otation for **U**nique description of **C**arbohydrate **S**equences) is used to establish an efficient interchange between the implemented databases, tools and applications. The service is available at: http://www.glycosciences.de.

1 Introduction

Carbohydrates are involved in a variety of fundamental biological processes, e.g. cellular differentiation, embryonic development or fertilisation. They are also involved in numerous pathological conditions as e.g. bacterial and viral infections, inflammatory diseases and cancer and therefore offer attractive pharmaceutical and diagnostic applications. In contrast to the genomic and proteomic area, no large data collections for carbohydrates have been compiled so far. The availability of such comprehensive databases, however,

* *glycosciences.de* has been supported by grants from the DFG (Deutsche Forschungsgemeinschaft). Dynamic Molecules project is promoted by the German Research Net (DFN) with means of the Federal Ministry for Research and Education (BMBF).

will be a prerequisite to successfully perform large-scale glycomics projects aiming to decipher new, so far unknown biological functions of glycans.

The structures of glycans as secondary gene products can not be easily predicted from the DNA sequence. Glycan sequences can not be described by a simple linear one letter code as each pair of monosaccharides can be linked in several ways and branched structures can be formed. Little of the bioinformatics algorithms developed for genomics/proteomics can be directly adapted for carbohydrates. Until recently there were only few bioinformatics databases and web applications dealing with glycobiology questions. The latest annually published list of molecular biology databases [Ga04] showed only three among about 300 databases dealing with glycobiology-related aspects.

However, the progressing glycomics projects will dramatically accelerate the understanding of the roles of carbohydrates in cell communication and hopefully lead to novel therapeutic approaches for treatment of human disease. The development of new and advanced bioinformatics tools, algorithms and data collections for glycobiology is an absolute requirement to successfully manage and analyze the large amount of data which will be produced by the upcoming glycomics project.

2 The *glycosciences.de* Portal

It is the aim of the *glycosciences.de* portal to provide a comprehensive platform, where glyco-related data collections originating from diverse open access resources as well as tools helping to interpret and analyse experimental data of complex carbohydrates are made available using standard internet technologies. Access is provided based either on (sub)-structural descriptions, bibliographic data, physicochemical properties and experimental data like NMR chemical shifts and MS-spectra.

2.1 Concept to link glyco-related data

The lack of generally accepted standards how to normalize glycan structures and exchange glycan formats hampers an efficient cross-linking and the automatic exchange of distributed data. Therefore, we have developed a unique notation for carbohydrate structures (LINUCS [BLLFvdL01]), that ensures a consistent logical interconnection between the different services. The *glycosciences.de* portal is prepared to be linked by external data collections which use the LINUCS-notation to describe carbohydrate structures. A special service is provided where external databases can access the LINUCS description using the SOAP protocol, thus enabling a direct connection to the *glycosciences.de* portal. Vice versa appropriate connections to external databases can be established by *glycosciences.de* through cross-linking of external databases in case access to LINCUS description is provided. A more detailed specification of the different databases and tools provided by *glycosciences.de* is given in die following sections.

3 Available tools on *glycosciences.de*

A wide range of tools and applications can be freely accessed on *glycosciences.de* as we support the idea of Open Access and publicly available software, and therefore all of our services can be used without charge by anyone (*http://www.glycosciences.de/tools*).

3.1 Tools supporting the interpretation and analysis of experimental data

3.1.1 Mass Spectrometry

The structural complexity and diversity of carbohydrates makes structure elucidation a challenging task despite the progress in experimental techniques. In recent years, MS has become the method of choice for high sensitive protein as well as glycan identification and characterization. The spectra of glycans can be complicated and difficult to interpret without reference data. Unfortunately, no libraries of suitably pure and homogeneous standards have so far been compiled.

GlycoFragment can be used to calculate and display the main fragments (B- and C-, Z- and Y-, A- and X-ions) of oligosaccharides that should occur in MS-spectra [LvdL03]. The extended ASCII character set as recommended by IUPAC is used to input the sequence of complex oligosaccharides. The main focus of GlycoFragment is to support the manual assignment of all peaks contained in mass spectra of complex carbohydrates.

GlycoSearchMS Recently, the GlycoFragment algorithm was used to create databases containing all theoretically possible fragments of about 5000 N- and 1200 O-glycans. Additionally, the masses of inner fragments - two independent glycosidic fragmentations or a single glycosidic and a cross-ring fragmentation are included. The GlycoSearchMS algorithm compares each peak of a measured MS-spectrum with the calculated fragments of all entries contained in the database. The number of matched peaks within a certain tolerance is used to compute a score by which the best matching spectra are ranked. For each matched experimental peak the structure of the associated fragment can be displayed. The reliability of results retrieved by GlycoSearchMS depends heavily on the comprehensiveness of the data collection searched. Since the database needs only theoretically calculated lists of fragments, the completion of missing structures will be relatively easy. This approach seems to be applicative for the rapid identification of known N- and O-glycans in high-throughput projects since the procedure is similar to routinely used approach for automatic peptide identification.

3.1.2 NMR

NMR techniques can lead to a full structural characterization of oligosaccharides including the monosaccharide stereochemistry, the anomeric configuration, the linkage type and the

complete sugar sequence. The implemented NMR tools allow retrieving NMR-spectra based on (sub)structural search, for atoms in a specific chemical environment and a spectral search where all peaks of the library search are compared to an experimental peak list. The number of matched peaks within a certain tolerance is used to compute a score by which the best matching library spectra are ranked. When comparing chemical shift values it is important that the reference data is measured at the same temperature and that the data are based on the same internal reference or one that can be correlated in a simple manner.

3.1.3 X-ray

pdb2linucs automatically extracts information of carbohydrates from PDB (Protein Data Bank [BKW$^+$77]) files and displays it using the LINUCS-notation. Many PDB entries contain carbohydrate structures, but the lack of a common standard nommenclature as it exists for amino acids complicates finding those carbohydrate information. Entire oligosaccahrides are sometimes encoded in one single residue. Information about carbohydrate linkages is often missing, and if present, it is not in a unique format and therfore difficult to find.

pdb-care aids experimentalists in detecting discrepancies in connectivities and nomenclature as they occur in about 30% of the carbohydrate-containing PDB entries as a recent study has revealed [LFvdL04]. The most common type of errors found in the PDB is a wrong assignment of the α-/β-isoforms. For example, there are two different PDB residue names for mannoses: MAN encodes for α-D-Man*p*, BMA for β-D-Man*p*. We have found 263 entries that contain at least one β-D-Man*p* named MAN. The opposite case, α-D-Man*p* named BMA, was found in 10 entries only.

3.2 Conformations of complex carbohydrates and their visualisation

SWEET2 is a program that rapidly converts the commonly used sequence information of complex carbohydrates directly into a preliminary but reliable 3D model. The basic idea is to link preconstructed 3D molecular templates of monosaccharides in a specific way of binding as defined in the sequence information. In a subsequent step a fast routine to explore the conformational space for each glycosidic linkage has been implemented. Systematic rotations around the glycosidic linkages are performed, calculating the van der Waals interactions for each step of rotation. The user interaction is supported by an input spreadsheet consisting of a grid of sugar symbol and connection type cells. Several ways to visualise and to output the generated structures and related information are implemented.

Dynamic Molecules is the first internet portal which provides interactive access to the techniques of molecular dynamics simulations via standard Web technologies and using only publicly available software. The ”expert mode”has been specially developed to explore the conformational space of oligosaccharides.

PDB2MultiGIF Visualization of chemical 3D structures on the web comes with problems because the web browser cannot display chemical structures without the help of additional software. If you create a page with a 3D structure of a molecule and the visitor of your page does not use this special viewer software for displaying molecules it cannot get the whole information of the page which should be meditated. PDB2MultiGIF takes the 3D structure and generates an animated image which can be displayed using any browser. Thus every visitor of your page can get the whole information.

4 Summary and outlook

Keeping scientific data collections up-to-date and feeding in new data continuously is one of the most time-consuming and thus expensive tasks. Therefore, new updating strategies for the GlycosciencesDB will be explored by the development and evaluation of automatic procedures to extract all public available data and facts for a specific glycan and try to classify, assign and annotate it according to predefined schemes and subjects. To guarantee high quality and data integrity of the data collections, the automatically retrieved and classified data will only enter the master data collection after human inspection and – if required – additional annotation and cross-links with other related information.

References

[BKW+77] Bernstein, F., Koetzle, T., Williams, G., Meyer, J. E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T., and Tasumi, M.: The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542. 1977.

[BLLFvdL01] Bohne-Lang, A., Lang, E., Forster, T., and von der Lieth, C.: Linucs: linear notation for unique description of carbohydrate sequences. *Carbohydr. Res.* 336(1):1–11. 2001.

[Ga04] Galperin, M.: The molecular biology database collection: 2004 update. *Nucl. Acids. Res.* 32:D3–D22. 2004.

[LFvdL04] Lütteke, T., Frank, M., and von der Lieth, C.: Data mining the protein data bank: Automatic detection and assignment of carbohydrate structures. *Carbohydrate Research.* 339:1015–1020. 2004.

[LvdL03] Lohmann, K. and von der Lieth, C.: GLYCO-FRAGMENT: A web tool to support the interpretation of mass spectra of complex carbohydrates. *Proteomics.* 3:2028–35. 2003.