Glycobiology vol. 21 no. 4 pp. 493–502, 2011
doi:10.1093/glycob/cwq188
Advance Access publication on November 23, 2010

EUROCarbDB: An open-access platform for glycoinformatics

Claus-Wilhelm von der Lieth[2,†], Ana Ardá Freire[3], Dennis Blank[4], Matthew P Campbell[5,6,11], Alessio Ceroni[7], David R Damerell[7], Anne Dell[7], Raymond A Dwek[6], Beat Ernst[8], Rasmus Fogh[9,12], Martin Frank[2], Hildegard Geyer[4], Rudolf Geyer[4], Mathew J Harrison[9,13], Kim Henrick[9,14], Stefan Herget[2], William E Hull[2], John Ionides[9,14], Hiren J Joshi[2,9], Johannis P Kamerling[3], Bas R Leeflang[3], Thomas Lütteke[3,12], Magnus Lundborg[10], Kai Maass[4,15], Anthony Merry[9], René Ranzinger[2,16], Jimmy Rosen[3], Louise Royle[5,6], Pauline M Rudd[5,6], Siegfried Schloissnig[2], Roland Stenutz[10], Wim F Vranken[9,14], Göran Widmalm[10], and Stuart M Haslam[1,7]

[2]Core Facility, Molecular Structure Analysis, German Cancer Research Center, Heidelberg, Germany; [3]Bijvoet-Center for Biomolecular Research, University of Utrecht, Utrecht, The Netherlands; [4]Institute of Biochemistry, Faculty of Medicine, Justus, Liebig University, Giessen, Germany; [5]Dublin-Oxford Glycobiology Laboratory, National Institute for Bioprocessing Research and Training (NIBRT), Conway Institute, University College Dublin, Dublin, Ireland; [6]Department of Biochemistry, Oxford Glycobiology Institute, University of Oxford, UK; [7]Division of Molecular Biosciences, Faculty of Natural Sciences, Biochemistry Building, Imperial College London, South Kensington Campus, London SW7 2AZ, UK; [8]Department of Pharmaceutical Science, University of Basel, Basel Switzerland; [9]European Bioinformatics Institute, Hinxton, UK; and [10]Organic Chemistry, Stockholm University, Stockholm, Sweden

Received on August 19, 2010; revised on November 3, 2010; accepted on November 3, 2010

The EUROCarbDB project is a design study for a technical framework, which provides sophisticated, freely accessible, open-source informatics tools and databases to support glycobiology and glycomic research. EUROCarbDB is a relational database containing glycan structures, their biological context and, when available, primary and interpreted analytical data from high-performance liquid chromatography, mass spectrometry and nuclear magnetic resonance experiments. Database content can be accessed via a web-based user interface. The database is complemented by a suite of glycoinformatics tools, specifically designed to assist the elucidation and submission of glycan structure and experimental data when used in conjunction with contemporary carbohydrate research workflows. All software tools and source code are licensed under the terms of the Lesser General Public License, and publicly contributed structures and data are freely accessible. The public test version of the web interface to the EUROCarbDB can be found at http://www.ebi.ac.uk/eurocarb.

Keywords: databases / glycoinformatics / informatics tools / open-source

Introduction

The inherent complexity of glycan structures and the microheterogeneity in naturally occurring glycans make analysis of glycoconjugates very challenging. A complete and detailed characterization of glycan structures is a time-consuming analytical process that typically requires a complex approach for detecting small quantities of glycans on low-abundance glycoproteins. It is widely accepted that glycosylation is the most complex and prevalent post-translational modification, which is important for protein folding, stability and function, with implications in signal transduction, differentiation, immunity, inflammation, cancer and other disease progressions (Rudd et al. 2001; Morelle and Michalski 2007; Arnold et al. 2008; Marth and Grewal 2008; Zaia 2008; An et al. 2009; Dennis et al. 2009; Taniguchi et al. 2009; Tharmalingam et al. 2010). Our understanding of glycan function is rapidly expanding through our improved understanding of the molecular glycosylation machinery and with advances in high-throughput glycomic strategies and glycoinformatics.

Over the past decade, there has been rapid progress in the development and availability of glycoanalytical technologies (Morelle and Michalski 2007; Widmalm 2007; Royle et al. 2008; Zaia 2008; Bielik and Zaia 2009; Blow 2009; North et al. 2009; Tharmalingam et al. 2010; Vanderschaeghe et al. 2010). These technical innovations and the increasing amounts of data generated by high-throughput techniques

[1]To whom correspondence should be addressed: Tel: +44-20-75945222; Fax: +44-20-72250458; e-mail: s.haslam@imperial.ac.uk
[†]Claus-Wilhelm von der Lieth deceased 16 November 2007.
[11]Present address: Department of Chemistry and Biomolecular Sciences, Biomelecular Frontiers Research Centre, Macquarie University, Sidney, Australia.
[12]Present address: Institute of Biochemistry und Endocrinology, Faculty of Veterinary Medicine, Justus, Liebig University, Giessen, Germany.
[13]Present address: Centenary Institute, Royal Prince Alfred Institute, Camperdown, Australia.
[14]Present address: Protein Data Bank in Europe (PDBe), EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
[15]Present address: Institute for Inorganic and Analytical Chemistry, Faculty of Biology and Chemistry, Justus Liebig University, Giessen, Germany.
[16]Present address: Complex Carbohydrate Research Center, University of Georgia, GA, USA.

allow for the rapid structural analysis of complex glycans. The complexity and volume of data routinely generated now necessitate bioinformatic solutions in the form of databases and analytical tools, either web-based or standalone, to support data interpretation and data storage (von der Lieth et al. 2004, 2006; Frank and Schloissnig 2010). Thus, new ideas are required for the development of universally accepted mechanisms and standards for storing both structural and experimental data and for improved data annotation and dissemination of the conclusions drawn from such experiments.

In contrast to the genomics and proteomics fields, the glycosciences lack accessible, curated and comprehensive data collections that summarize the structure, characteristics, biological origin and potential function of glycans which have been experimentally verified and reported in the literature. Over the past two decades, efforts in the development of bioinformatic resources have increased, including projects by international consortia (Table I in Frank and Schloissnig 2010) to develop databases that support the growing requirement for storing and presenting data to address the needs of the glycobiology community.

A majority of these databases have been derived, in part, from the first international effort to set up a carbohydrate database, the Complex Carbohydrate Structure Database (CCSD), commonly referred to as CarbBank, which was developed by the Complex Carbohydrate Research Center (Doubet et al. 1989; Doubet and Albersheim 1992). The project started in the late 1980s but ceased in 1997 due to lack of funding support. The final database contained ~50,000 records comprising 23,000 unique sequences in the CarbBank notation with associated biological background, chemical data and publication information. This data set has been used by recent initiatives to seed new databases with a basic set of glycan structure records.

Three prominent initiatives are the Kyoto Encyclopedia of Genes and Genomes (KEGG) Glycan (Aoki et al. 2004; Hashimoto et al. 2006), the US consortium for functional glycomics (CFG; Raman et al. 2006) and the GLYCOSCIENCES.de portal (Lütteke et al. 2006). Briefly, KEGG Glycan is an integrated knowledge base of protein networks with genomic and chemical information and provides access to sugar structures through the manually drawn pathway maps representing the current knowledge of glycan biosynthesis and metabolism for various species. The CFG databases have been designed to host and integrate experimental data produced by the various core groups of the consortium. The GLYCOSCIENCES.de portal is a resource for linking glycan-related data originating from various resources through a unique structure description. Special emphasis has been made to provide easy access to the available experimentally determined spatial structures of glycans and to analyze their interactions with proteins. The CFG and GLYCOSCIENCES.de databases contain collections of experimental data from mass spectrometry (MS) profiles and nuclear magnetic resonance (NMR) experiments.

Although many carbohydrate databases have been implemented (Cooper et al. 2001, 2003; Campbell et al. 2008), the desirable features for structure-based cross-linking and automated exchange of data between established resources are lacking. This weakness arises from the use of different database-specific structure encoding formats that are difficult or impossible to cross-translate in a facile manner due to semantic and syntactic nuances that are intractable to automatic translation (Doubet and Albersheim 1992; Bohne-Lang et al. 2001; Aoki et al. 2004; Kikuchi et al. 2005; Campbell et al. 2008). Thus, we are faced with a variety of incomplete databases characterized by disconnected and incompatible collections of experimental and structural data. This situation not only hinders the development of bioinformatic tools but also the realization of platforms for large-scale glycomics and glycoproteomics studies. As glycan-related databases improve in coverage and quality, it is of growing interest to consider criteria and solutions that maximize the value that can be extracted. Recent reviews highlight the approaches and difficulties the glycoinformatics community is facing and those steps being taken to create well-curated and annotated databases and efficient tools (Aoki-Kinoshita 2008; Packer et al. 2008).

Here, we present EUROCarbDB (http://www.eurocarbdb.org), an Infrastructure Design Study funded under the sixth EU framework program, to establish the technical requirements for developing a centralized and standardized database architecture for carbohydrate-related data (structure and analytical data). The study focused on the evaluation and development of a robust infrastructure that supports the established analytical methods. The initial implementation of the EUROCarbDB outlined here includes workflows and tools developed specifically for the key analytical techniques of high-performance liquid chromatography (HPLC), MS and NMR spectroscopy. The platform offers the following features: (i) an introduction and/or recommendation of formats and nomenclatures for encoding carbohydrate structures; (ii) relational database schemas for storing curated structural and experimental data together with a web-based user interface designed to populate and curate that data; (iii) various web-based tools for visualizing and querying all stored information and (iv) bioinformatics research tools designed to expedite aspects of glycan analysis and structural elucidation. Lastly, software libraries and bioinformatics tools produced by the project are available at http://eurocarb.googlecode.com, under the terms of the Lesser General Public License (http://www.gnu.org/licenses/lgpl.html).

Results and discussion

The design goals and concepts of the EUROCarbDB are far-reaching and by no means limited to the resources presented here. A specific design objective of the architecture of the database was to allow for the extension and incorporation of new modules and tools to support further types of experimental data and workflows. In this fashion, the emphasis throughout the project has been to develop a comprehensive framework that promotes accessibility and longevity through a modular design approach and by adopting an open-source philosophy. It is our hope that this will in turn drive the continued growth and promotion of unified glycoinformatic tools analogous to those that have facilitated the rapid uptake of proteomics and genomics and will allow future integration

with other complementary database resources such as the PDBe: Protein Data Bank in Europe (Velankar et al. 2010).

Glycan structure encoding

A review of current and discontinued database projects identified that a major issue in the design of a robust and future-oriented data resource is the availability of an infrastructure which can conveniently handle structural data at both the internal database and user levels. The main goal of these developments was to simplify the manual input of structures while providing the capabilities for encoding all structural characteristics of a glycan sequence in a unique and computer-readable form. Note that throughout this report the term "structure" will be used to refer to the more or less detailed constitution, sequence (linkage) and stereochemistry information available for a particular glycan and will generally not be used to denote a particular 3D spatial structure (atomic coordinates) in the molecular modeling sense.

To motivate scientists to use newly designed resources, we developed a rapid and intuitive glycan structure editor (see the *GlycanBuilder* section) that is embedded with a controlled set of precise monosaccharide descriptions (see the *MonosaccharideDB* section) for the rapid generation of glycan structures. The analysis of existing structures and formats indicated that a new, more comprehensive format (Table I) was required for the accurate encoding of structural details in a form that can be read by the human user, as well as by computer algorithms, and can serve as a unique identifier for database entries.

GlycoCT. The analysis of previous encoding formats (Table I) showed that none of them is capable of dealing with the full complexity to be expected for experimentally derived carbohydrate sequence data. Therefore, we developed a new encoding scheme for complex carbohydrates, named GlycoCT (Herget et al. 2008).

This format utilizes a connection table (CT) approach, instead of a linear encoding scheme, and features a controlled vocabulary for monosaccharide building blocks and substituents, according to the IUPAC nomenclature. GlycoCT uses a block concept to describe residues, linkages and frequently occurring special features, such as repeating units. The implementation of sorting rules and the ability to encode ambiguous or only partially defined structures (e.g. uncertain terminal residue positions) as well as fully resolved structures assures the uniqueness of the GlycoCT signature for each given structure. It exists in two variants, a condensed form and a more verbose XML syntax, which are described in detail elsewhere (Herget et al. 2008) and http://www.eurocarbdb.org/recommendations/encoding. The main intent of GlycoCT is to be used as a unique encoding of a glycan structure that may be readily translated to other existing text formats and graphical representations, such as Oxford University format, CFG notation and IUPAC text format.

Structure translation. As a seed for initial database development, 50,000 entries were imported from the public CarbBank repository to EUROCarbDB via an internally developed routine. After removal of any aglyca (e.g. amino acids or lipids, which are part of the sequence in CarbBank), the remaining carbohydrate moiety was exported into the GlycoCT sequence format and subsequently stored in EUROCarbDB.

GlycanBuilder. Given the nonlinear and frequently complex nature of carbohydrate sequences, many researchers prefer to visualize glycans graphically, rather than as textual sequence strings. The GlycanBuilder is an embeddable software tool (Java applet), which can be used to visually compose glycan structures from a set of monosaccharide building blocks (Ceroni et al. 2007; http://www.glycoworkbench.org/wiki/GlycanBuilder). The "view" menu of the GlycanBuilder tool allows sequences to be generated and viewed in any of the five graphical display schemes illustrated in Figure 1. The user can rapidly specify a glycan structure by simply selecting successive residues and their points of attachment. The growing structure is displayed using one of the selectable symbolic notations. The list of structural constituents comprises an exhaustive collection of saccharides, substituents, reducing-end markers and saccharide modifications. All of the stereochemical information about a saccharide, e.g. anomeric configuration, chirality, ring configuration and linkage positions, can be specified. The GlycanBuilder is used in the

Table I. Comparison of carbohydrate structure encoding formats in terms of structural features, which can be encoded in an unrestricted manner (+), restricted manner (o) or not at all (–).

	GlycoCT	KCF	LINUCS	CarbBank format	BCSDB format	GlycoBase (Dublin)	Linear Code®	CabosML	Glycosuite format
Encodable features									
Unknown linkage positions	+	+	+	+	+	+	+	–	+
Repeat units	+	+	+	+	0	–	+	+	+
Cyclic sugars	+	–	+	+	–	–	+	+	–
Undeterminded terminal residues	+	–	–	+	–	+	+	–	0
Alternative residues	+	–	–	–	+	–	+	–	+
Nonstoichometric modifications	+	–	–	+	+	–	–	–	0
Compositions	+	+	–	–	–	–	–	–	–
Unique sequences	+	–	0	–	+	+	+	–	+
Multiple connections	+	+	–	–	+	–	–	–	–

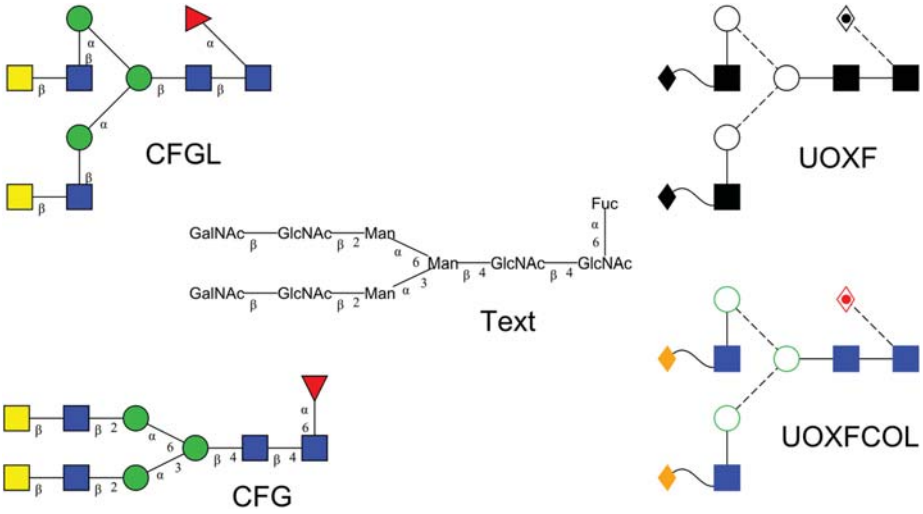

Fig. 1. Carbohydrate representations used in GlycanBuilder. Users can select from five graphical display schemes for glycan structures. As an example structure a complex *N*-glycan is shown in the IUPAC Text mode, the CFG symbolic format with linkage labels, the CFGL format with linkage positions shown geometrically and the Oxford black & white (UOXF) and color (UFOXCOL) schemes, where linkage positions are shown by geometry and anomeric configurations are denoted by dashed (α) or solid (β) lines. Note that this structure is indefinite since the linkage positions of the terminal GalNAc residues are not defined.

web interface of the EUROCarbDB as the standard input facility for structures in all *contribute*, *browse* and *search* pages. Furthermore, the GlycanBuilder is embedded into the GlycoWorkbench software (Ceroni et al. 2008) as a component for drawing structure candidates in the process of annotating MS spectra (see the *Software tools* section below). Herein, sequences can be exported in GlycoCT as an XML file or in any of several standard text or graphical formats. Finally, following the open-source character of the EUROCarbDB project, the GlycanBuilder was used in other database projects (see http://www.glycome-db.org) as a structure input facility for search functions (Ranzinger et al. 2008, 2009).

MonosaccharideDB. A major problem with previous carbohydrate databases has been the frequent use of nonuniform residue names (aliases or synonyms) for a given monosaccharide, e.g. Fuc/6-deoxy-Gal or 2-deoxy-Glc/2-deoxy-Man/2-deoxy-araHex. This problem can be solved by using a controlled vocabulary of unique monosaccharide names. In view of the large number of known monosaccharides and the even larger number theoretically feasible, it is impractical to adequately define a static dictionary with all possible monosaccharides and synonyms. Therefore, in EUROCarbDB, the dynamic, self-extensible dictionary and routines of MonosaccharideDB are implemented to ensure residue name consistency. The dictionary can also be accessed independently at http://www.monosaccharidedb.org.

In addition to its own internal format (closely related to GlycoCT), MonosaccharideDB currently supports six different notation schemes for carbohydrates: GlycoCT (Herget et al. 2008); the LinearCode® used by the CFG (Raman et al. 2006) and Glycominds Ltd.; the LINUCS notation (Bohne-Lang et al. 2001) used in GLYCOSCIENCES.de (Lütteke et al. 2006); the schemes of the CCSD (Doubet et al. 1989; the extended IUPAC notation of CarbBank); the Protein Data Bank (PDB) (Henrick et al. 2008); and the Bacterial Carbohydrate Structure Database (BCSDB; Toukach and Knirel 2005). To ensure unambiguous encoding of glycans, MonosaccharideDB is capable of translating carbohydrate residue names from different encoding schema into a unique notation and vice versa.

EUROCarbDB database

The central aim of the EUROCarbDB design study was to develop a comprehensive, relational database of experimentally determined carbohydrate structures, supported by biological context information, literature references and empirical structural evidence. For the purpose of the initial design study, MS, HPLC and NMR data were designed by the EUROCarbDB. The interaction of these database sections is summarized in Figure 2. Here, we outline the central glycan structure domain as well as the three evidence domains (and the experimental features they support). Technical details will be described in subsequent publications.

Core database. The purpose of the core database is to act as a central resource of glycan structures, associated biological contexts and references. Glycan structures are considered to be the focal point of the core database and they are used to connect the evidence domains to the core domain (Figure 2). The initial version of the EUROCarbDB was seeded with glycan structures as described in the *Structure translation* section. Additional structures and data have also been incorporated by EUROCarbDB partners, for example, from the GlycoBase database (Campbell et al. 2008). Each structure entry in EUROCarbDB is encoded in the database in GlycoCT format as described previously.

The association of structure to biological source is captured by the shared biological context module. A biological context

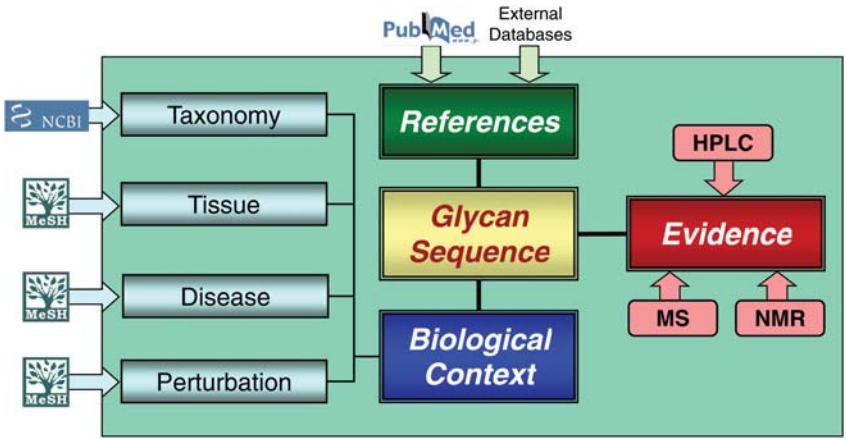

Fig. 2. Architecture of EUROCarbDB. The core database contains the four components structure, biological context, (experimental) evidence and references. Biological context comprises several categories (blue) which utilize the controlled vocabularies obtained from external databases (NCBI, MeSH). References (green) are also obtained from external databases such as PubMed. A key feature of EUROCarbDB is the internal storage of experimental evidence (right) for a glycan sequence.

within the EUROCarbDB is defined as the amalgamation of a specific taxonomy and tissue, together with a varying number of disease and perturbation associations. EUROCarbDB exclusively uses controlled vocabularies derived from and linked directly to the public domain ontologies of NCBI Taxonomy (Sayers et al. 2009), and the MeSH categories A, C and D ("Anatomy", "Diseases" and "Chemicals and Drugs" respectively; Lowe and Barnett 1994). These controlled vocabularies are also innately hierarchical, such that a structure linked to "*Homo sapiens*", "Hepatocytes" and "Hepatic cellular carcinoma" is hierarchically related to and therefore searchable by reference to the more general nodes for "Mammalia", "Liver" and "Neoplasms".

EUROCarbDB can store references for both glycan structures and experimental evidence in the form of PubMed IDs, external database identifiers and other internet resources. The current EUROCarbDB policy for submission of a unique glycan sequence requires that at least one published literature reference be supplied which describes the structure elucidation process, a link to an external database or that experimental evidence be submitted.

The ability to associate any structure entry with corresponding properties or features results in a fully annotated database of sequence information and metadata collections, which in turn provides the technical basis for sophisticated querying the database via the user interface.

Experimental evidence. A key design objective of the EUROCarbDB was the creation of a database infrastructure which supports the deposition, analysis, annotation and curation of experimental data from the analytical techniques of HPLC, MS and NMR—the most popular techniques for the structure elucidation of carbohydrates. The experimental data section of EUROCarbDB currently supports the input (archival) of raw data, and the interpretation of these data can provide the experimental evidence for one or more carbohydrate structures in a given biological context. However, EUROCarbDB can also accept new structures for which the evidence (raw or interpreted data) is no longer available. In this case, at least one published article in a peer-reviewed journal or the link to an external database entry must be provided.

High-performance liquid chromatography. One major aim of the EUROCarbDB design study was to develop and provide a framework for algorithms and tools to facilitate the deposition and interpretation of HPLC data. Here, an innovative glycoinformatic suite of well-curated databases and analytical tools has been developed to support the growing demand for HPLC data processing resources, such as GlycoBase and auto-glucose unit (GU; Campbell et al. 2008). These tools are complemented by a series of data entry and processing workflows supported by a robust database framework with interfaces for retrieving structural and experimental data.

The data model is composed of one experimental evidence section and one metadata section—allowing for the complete description of an HPLC experiment. The experimental evidence section is used to associate glycan sequences (stored within the central EUROCarbDB sequence database) with their HPLC elution positions (normalized retention times, expressed as GU values). The metadata section allows for the storage of the following information: *instrument type*, *chromatographic conditions* and *data processing*. In addition, the associated biological context metadata can be stored for each HPLC experiment—using the previously defined facilities of the core component of the EUROCarbDB.

Mass spectrometry. Is a widely used experimental technique for the elucidation of glycan structures, either as found attached to specific entities or within complex samples (Harvey 1999, 2006, 2008). Therefore, an extensive collection of tools and databases has been created.

The MS data model and interface are completely integrated within the facilities of the EUROCarbDB core database and were designed to provide comprehensive workflows from the acquisition of raw MS data to the final deposition of data into

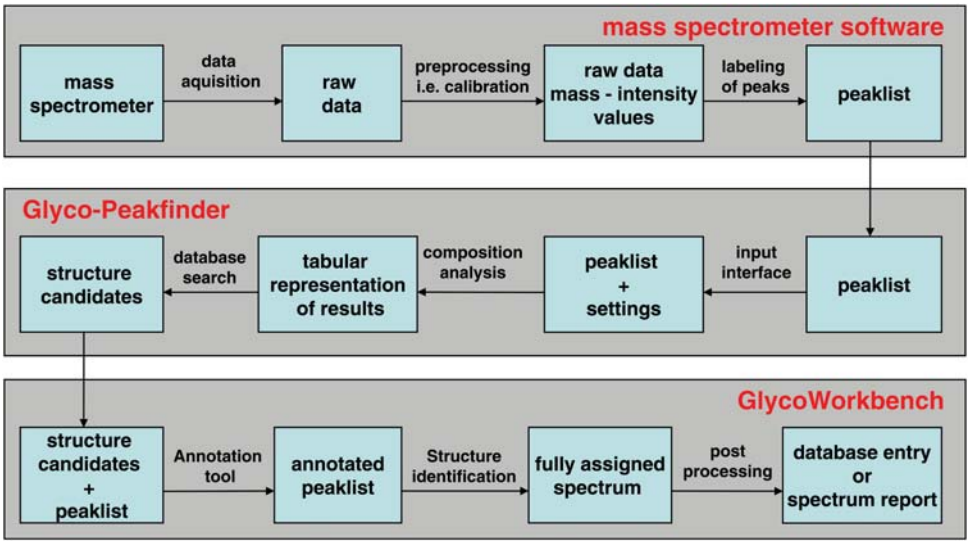

Fig. 3. Proposed experimental workflow for uploading MS data into EUROCarbDB. Starting from acquired raw data, the subsequent use of MS vendor's software (top row) and the newly developed tools Glyco-Peakfinder (middle row) and GlycoWorkbench (bottom row) allows the user to perform structure determination, peak assignments and annotation and data submission to EUROCarbDB.

the database (Figure 3). In combination with the tools, described in the software tools section, the workflows allow experimentalists to cascade all steps of structure identification. Starting with the generation of peak lists (using existing MS software), the new tool called Glyco-Peakfinder is used to calculate monosaccharide compositions from which structure candidates are generated. The GlycoWorkbench tool is then used to match the peak lists with the structure candidates—producing fully annotated spectra. Finally, the web interface of the EUROCarbDB (discussed subsequently) is used to upload the results into the MS evidence domain, which has been designed to store raw spectra, their subsequent annotations and the metadata required to fully describe each experiment, including biological context, experimental conditions and other parameters.

Nuclear magnetic resonance. The inherent volume and complexity of the raw and interpreted NMR data generated for glycan structure elucidation requires a sophisticated data model. After the evaluation of various alternatives, it was decided to use the CCPN data model (Collaborative Computing Project for NMR, http://www.ccpn.ac.uk/index.html), for storing NMR evidence in EUROCarbDB (Vranken et al. 2005). At the moment only complete CCPN projects can be submitted to the database, but the manual input of experimental data and assignments would be easy to implement.

The EUROCarbDB web portal (www.ebi.ac.uk/eurocarb). The web-based user interface of the EUROCarbDB allows for searching and browsing deposited information as well as for submitting new structures and experimental data to the database. Users can browse the database by various aspects of structure, experimental evidence, taxonomy, tissue, disease and perturbation. Each glycan structure has an individual "entry" page (Figure 4) that can be accessed either by its EUROCarbDB identifier or via links from browsing and querying pages.

Glycan structures may be queried by a variety of structural, reference and biological context criteria. Queries may also be cascaded, in which results from previous queries may be narrowed further by a subsequent query. EUROCarbDB also performs substructure queries entirely within the database via a novel algorithm, which allows for very fast searches.

Users initiate queries of the database using the top navigation bar that provides links for: browse, search, contribute and tools. Below the navigation bar is the main content area that is used for, query input and result set display. Panels on the right-hand side of the interface allows for XML export and the selection of glycan structure visual representation. This component displays links to other databases and components of the EUROCarbDB that are related to the currently displayed information. For example, the results of an automatic substructure search are displayed on this component when a glycan information page is being viewed.

For the upload of structures and data to the EUROCarbDB, contributors are guided through a sequence of input pages. Beginning with drawing the glycan structure, followed by selecting the biological context, entering the reference data and, finally, uploading spectra and/or chromatograms in a step-by-step fashion. The GlycanBuilder tool is also used extensively throughout the user interface as an input mechanism wherever structural inputs are required.

Software tools

From the beginning, it was obvious that the development of tools for semi-automatic interpretation of MS spectra, NMR data or HPLC profiles was fundamental to the aims of the project. Therefore, the software tools Glyco-Peakfinder and GlycoWorkbench for the interpretation of MS data, autoGU for HPLC analysis and ProSpectND and CASPER (Computer-Assisted SPectrum Evaluation of Regular polysaccharides) for NMR analysis were produced. Figure 5 displays the relationships that exist between these tools and the EUROCarbDB.

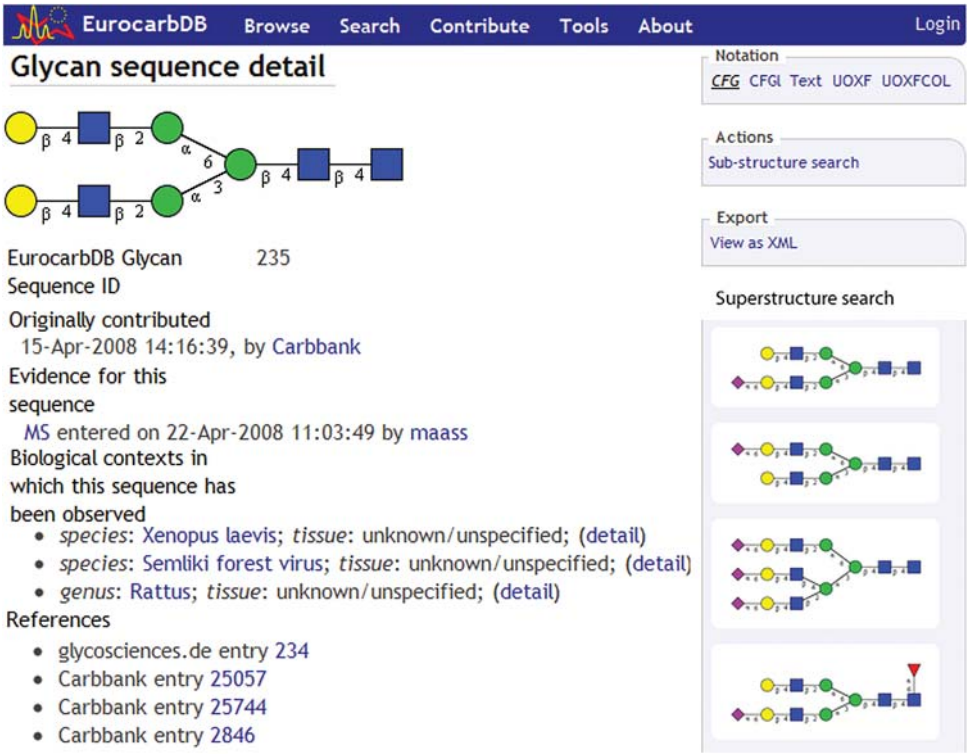

Fig. 4. EUROCarbDB web portal. Screenshot demonstrates the information shown for each glycan structure within the database. Features of note include: links to evidence, annotated biological contexts and references, control panel to change graphical representation of structures, automatic superstructure search and main navigation bar.

GlycoBase and autoGU. GlycoBase (http://glycobase.nibrt.ie:8080/database/show_glycobase.action) is an evidence database containing the normalized HPLC elution positions for 2-aminobenzamide-labelled glycan structures. A standalone release of the database was described previously (Campbell et al. 2008), which has now been extended and fully integrated within the EUROCarbDB framework. Briefly, GlycoBase contains GU values for defined glycans obtained from a variety of sources. The information concerning biological source, exoglycosidase digestions (enzymes used, glycan products) and supporting publications is available through the EUROCarbDB core database model. All structures were determined by a combination of normal-phase HPLC with exoglycosidase sequencing and/or were confirmed by MS, e.g. MALDI-MS, ESI-MS, ESI-MS/MS, LC-ESI-MS and LC-ESI-MS/MS.

An intuitive workflow allows users to upload a series of retention values and percentage areas from propriety HPLC software. When combined with exoglycosidase information using a simple step-by-step workflow a GlycoBase search (all data or a specified subset) can be performed. Initial putative structure assignments for the undigested retention list are progressively analyzed and subsequently refined with supporting exoglycosidase data, using the autoGU tool. Using data from a series of exoglycosidases, autoGU creates a refined list of structural assignments that match the supporting digest footprint, i.e. observed shifts in GU values are due to the cleavage of terminal monosaccharides whose identities depend on enzyme specificity. GlycoBase in combination with autoGU can be used to interpret and assign HPLC profiles; providing an invaluable tool that is frequently used by international biotechnology and pharmaceutical companies, as well as academic institutions (Saldova et al. 2007; Abd Hamid et al.

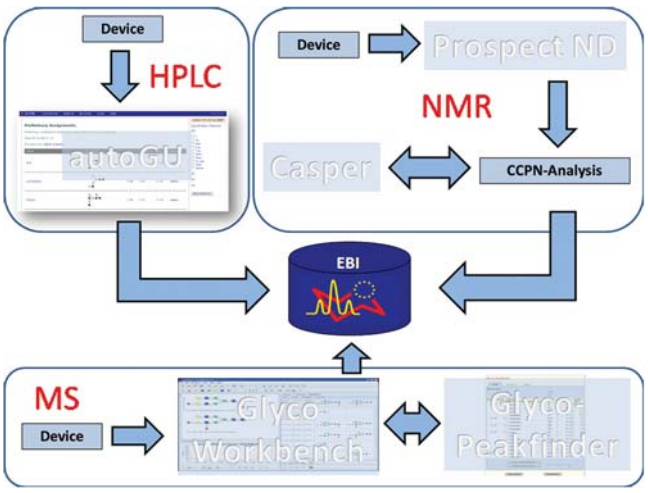

Fig. 5. Relationships between analysis tools and the EUROCarbDB. The figure represents the analytical workflows designed for EUROCarbDB, specifically highlighting the central role of the developed tools.

2008; Royle et al. 2008; Liu et al. 2010; Marino et al. 2010; Tharmalingam et al. 2010).

Glyco-Peakfinder. Glyco-Peakfinder (http://www.glyco-peakfinder.org; Maass et al. 2007) is a web-based software tool for semi-automatic de novo composition analysis of glycans, providing computer assistance in the analysis of fragmentation patterns and the laborious manual annotation of all kinds of carbohydrate MS spectra. Thus, the software can be used for the prediction and verification of the annotations which can be entered into EUROCarbDB, and workflows have been defined which integrate Glyco-Peakfinder into the input strategy for MS evidence in EUROCarbDB. The software functionalities are also available within GlycoWorkbench.

GlycoWorkbench. GlycoWorkbench (http://www.glycoworkbench.org; Ceroni et al. 2008) is a standalone desktop application that provides a suite of tools for the routine annotation and computer-assisted interpretation of glycan mass spectra. A key feature is the semi-automated correlation of observed fragment masses with computer-generated fragmentation patterns on input structures proposed by the researcher. The graphical interface provides an environment in which structure models can be rapidly assembled using GlycanBuilder, automatically matched with MS^n data and compared with assess the best candidate. GlycoWorkbench provides a fully integrated workflow to generate an annotated list of peaks from an MS experiment and to upload the results to the database. The software supports a wide variety of structural constituents and annotation options, as well as an exhaustive collection of fragmentation types.

Glyco-Peakfinder and GlycoWorkbench software tools have been widely used for calculation of monosaccharide compositions, annotation of mass spectra and identification of oligosaccharide fragments obtained by tandem MS. Examples are, inter alia, the structural characterization of *N*-glycans from the freshwater snail *Biomphalaria glabrata* (Lehr et al. 2007), glycomic analyses of *N*- and *O*-linked carbohydrate epitopes of the neural cell adhesion molecule CD24 from mouse brain (Bleckmann, Geyer, Lieberoth, et al. 2009; Bleckmann, Geyer, Reinhold, et al. 2009), the characterization of the acidic *N*-glycans of the zona pellucida of prepuberal pigs (von Witzendorff et al. 2009), glycomic analyses of human neutrophil *N*- and *O*-linked carbohydrate epitopes from Babu et al. (2009), analysis of the human seminal plasma glycome (Pang et al. 2009) and structural elucidation of glycosaminoglycans (Tissot et al. 2008).

Computer-assisted spectrum evaluation of regular polysaccharides

CASPER (http://www.casper.organ.su.se/eurocarbdb/casper.action) is a software tool which uses empirical increment rules based on experimental reference data to perform ^1H and ^{13}C NMR *chemical shift prediction* for oligo- and polysaccharides (Jansson et al. 2006; Stenutz 2009). The reference database has been derived from standardized measurements on a representative collection of mono-, di- and trisaccharides (substructures). The data for the constituent monosaccharides comprising any glycan are used to create a zero-order prediction set. Differences between the observed chemical shifts for mono- and disaccharides provide a set of first-order chemical shift increments (glycosylation shifts), whereas trisaccharides provide additional second-order chemical shift corrections for vicinal branch points. This means that CASPER can be used both as a reference tool to retrieve chemical shifts of monosaccharides as well as to calculate the chemical shifts of complex oligo- and polysaccharides for comparison with experimental data and thereby aid in structure assignment.

In a complementary manner, CASPER can be used for glycan structure prediction on the basis of NMR data. As input CASPER accepts uncorrelated ^1H and ^{13}C chemical shifts, or correlated pairs of $^{13}C/^1H$ chemical shifts (e.g. from 2D-NMR experiments). In addition, information about coupling constants at the anomeric centers ($^3J_{HH}$ and $^1J_{CH}$) can be entered to limit the number of possible structures and shorten the calculation times. CASPER can deal with uncertain linkage positions and unknown residue identities for cases where chemical analysis is incomplete, but the calculations will require more time. First, the software generates all possible structures consistent with data provided by the user. The chemical shifts for each structure are then predicted as described above and compared with the experimental chemical shifts of the unknown glycan. The structures are ranked according to the quality of agreement between observed and predicted chemical shifts. When necessary, the user can then more efficiently plan additional experiments for unambiguous structure confirmation.

CASPER is also used directly by the EUROCarbDB to check the consistency of new data during submission to the NMR evidence database and can warn the user about possible mistakes in NMR signal assignments (large deviations between observed and predicted chemical shifts). CASPER can use CCPN projects as data input and can also save results in that format.

ProSpectND. ProSpectND is an open-source, end-user software for the processing and evaluation of one- and multidimensional NMR data (http://sourceforge.net/projects/prospectnd/). It runs natively on all major platforms (Unix/X-windows, Microsoft Windows and Mac OSX). Through versatile scripting tools, raw data from both Bruker and Varian NMR spectrometers can be handled. ProSpectND provides comprehensive features for the entire data processing workflow. The optimally processed spectra can then be transferred to the CCPN analysis tools for the signal assignment phase. In addition, ProSpectND can be used to carry out exact quantum-mechanical one-dimensional ^1H-NMR spectrum simulations for defined spin systems. Such simulations, which include the second-order effects of strong coupling, have been shown to be of great assistance in confirming the details of carbohydrate signal assignments, especially in the case of overlapping multiplets. ProSpectND is feature rich and open source, making it a good alternative to commercial processing software.

Outlook

Recent experience clearly indicates that, following in the footsteps of genomics and proteomics, the next explosively developing fields in the biosciences will be glycomics and

metabolomics. The EUROCarbDB initiative has removed one of the key stumbling blocks impeding progress in glycomics by providing the glycobiology community with (i) universal standards for the representation of monosaccharides and complex glycans, (ii) a freely accessible database of known glycan structures and experimental evidence, which embodies these standards, (iii) freely accessible analytical tools for carbohydrate researchers and (iv) a technical framework of open-source code libraries and applications for carbohydrate bioinformatics.

Funding

The EUROCarbDB initiative (http://www.eurocarbdb.org, Design Studies Related to the Development of Distributed, Web-based European Carbohydrate Databases) was funded by the European Union as a Research Infrastructure Design Study implemented as a Specific Support Action under the FP6 Research Framework Program (RIDS Contract number 011952). D.R.D. is funded by the Biotechnology and Biological Sciences Research Council (BBF0083091 to A.D., S.M.H.).

Acknowledgements

We also thank the many developers who have contributed time and effort to produce the multitude of open-source code libraries used by EUROCarbDB.

Conflict of interest

None declared.

Abbreviations

CASPER, computer-assisted spectrum evaluation of regular polysaccharides; CCPN, Collaborative Computing Project for NMR; CCSD, Complex Carbohydrate Structure Database; CFG, consortium for functional glycomics; CT, connection table; GU, glucose unit; HPLC, high-performance liquid chromatography; KEGG, Kyoto Encyclopedia of Genes and Genomes; LGPL, Lesser General Public License; MS, mass spectrometry; NMR, nuclear magnetic resonance.

References

Abd Hamid UM, Royle L, Saldova R, Radcliffe CM, Harvey DJ, Storr SJ, Pardo M, Antrobus R, Chapman CJ, Zitzmann N, et al. 2008. A strategy to reveal potential glycan markers from serum glycoproteins associated with breast cancer progression. *Glycobiology*. 18:1105–1118.

An HJ, Kronewitter SR, de Leoz ML, Lebrilla CB. 2009. Glycomics and disease markers. *Curr Opin Chem Biol*. 13:601–607.

Aoki-Kinoshita KF. 2008. An introduction to bioinformatics for glycomics research. *PLoS Comput Biol*. 4:e1000075.

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.

Arnold JN, Saldova R, Hamid UM, Rudd PM. 2008. Evaluation of the serum *N*-linked glycome for the diagnosis of cancer and chronic inflammation. *Proteomics*. 8:3284–3293.

Babu P, North SJ, Jang-Lee J, Chalabi S, Mackerness K, Stowell SR, Cummings RD, Rankin S, Dell A, Haslam SM. 2009. Structural characterisation of neutrophil glycans by ultra sensitive mass spectrometric glycomics methodology. *Glycoconj J*. 26:975–986.

Bielik AM, Zaia J. 2009. Historical overview of glycoanalysis. *Methods Mol Biol*. 600:9–30.

Bleckmann C, Geyer H, Lieberoth A, Splittstoesser F, Liu Y, Feizi T, Schachner M, Kleene R, Reinhold V, Geyer R. 2009. O-glycosylation pattern of CD24 from mouse brain. *Biol Chem*. 390:627–645.

Bleckmann C, Geyer H, Reinhold V, Lieberoth A, Schachner M, Kleene R, Geyer R. 2009. Glycomic analysis of N-linked carbohydrate epitopes from CD24 of mouse brain. *J Proteome Res*. 8:567–582.

Blow N. 2009. Glycobiology: A spoonful of sugar. *Nature*. 457:617–620.

Bohne-Lang A, Lang E, Forster T, von der Lieth CW. 2001. LINUCS: Linear notation for unique description of carbohydrate sequences. *Carbohydr Res*. 336:1–11.

Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM. 2008. GlycoBase and autoGU: Tools for HPLC-based glycan analysis. *Bioinformatics*. 24:1214–1216.

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*. 7:2–3.

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.

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.

Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH. 2003. GlycoSuiteDB: A curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res*. 31:511–513.

Dennis JW, Lau KS, Demetriou M, Nabi IR. 2009. Adaptive regulation at the cell surface by *N*-glycosylation. *Traffic*. 10:1569–1578.

Doubet S, Albersheim P. 1992. CarbBank. *Glycobiology*. 2:505.

Doubet S, Bock K, Smith D, Darvill A, Albersheim P. 1989. The Complex Carbohydrate Structure Database. *Trends Biochem Sci*. 14:475–477.

Frank M, Schloissnig S. 2010. Bioinformatics and molecular modeling in glycobiology. *Cell Mol Life Sci*. 67:2749–2772.

Harvey DJ. 1999. Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrom Rev*. 18:349–450.

Harvey DJ. 2006. Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update covering the period 1999–2000. *Mass Spectrom Rev*. 25:595–662.

Harvey DJ. 2008. Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update covering the period 2001–2002. *Mass Spectrom Rev*. 27:125–201.

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.

Henrick K, Feng Z, Bluhm WF, Dimitropoulos D, Doreleijers JF, Dutta S, Flippen-Anderson JL, Ionides J, Kamada C, Krissinel E, et al. 2008. Remediation of the protein data bank archive. *Nucleic Acids Res*. 36: D426–D433.

Herget S, Ranzinger R, Maass K, von der Lieth CW. 2008. GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr Res*. 232:2162–2171.

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.

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.

Lehr T, Geyer H, Maass K, Doenhoff MJ, Geyer R. 2007. Structural characterization of *N*-glycans from the freshwater snail *Biomphalaria glabrata* cross-reacting with *Schistosoma mansoni* glycoconjugates. *Glycobiology*. 17:82–103.

Liu L, Telford JE, Knezevic A, Rudd PM. 2010. High-throughput glycoanalytical technology for systems glycobiology. *Biochem Soc Trans*. 38:1374–1377.

Lowe HJ, Barnett GO. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA*. 271:1103–1108.

Lütteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, von der Lieth CW. 2006. GLYCOSCIENCES.de: An Internet portal to support glycomics and glycobiology research. *Glycobiology*. 16:71R–81R.

Maass K, Ranzinger R, Geyer H, von der Lieth CW, Geyer R. 2007. Glyco-Peakfinder – *de novo* composition analysis of glycoconjugates. *Proteomics*. 7:4435–4444.

Marino K, Bones J, Kattla JJ, Rudd PM. 2010. A systematic approach to protein glycosylation analysis: A path through the maze. *Nat Chem Biol*. 6:713–723.

Marth JD, Grewal PK. 2008. Mammalian glycosylation in immunity. *Nat Rev Immunol*. 8:874–887.

Morelle W, Michalski JC. 2007. Analysis of protein glycosylation by mass spectrometry. *Nat Protoc*. 2:1585–1602.

North SJ, Hitchen PG, Haslam SM, Dell A. 2009. Mass spectrometry in the analysis of *N*-linked and *O*-linked glycans. *Curr Opin Struct Biol*. 19:498–506.

Packer NH, von der Lieth CW, 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.

Pang PC, Tissot B, Drobnis EZ, Morris HR, Dell A, Clark GF. 2009. Analysis of the human seminal plasma glycome reveals the presence of immunomodulatory carbohydrate functional groups. *J Proteome Res*. 8:4906–4915.

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.

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.

Ranzinger R, Herget S, Wetter T, von der Lieth CW. 2008. *GlycomeDB* - integration of open-access carbohydrate structure databases. *BMC Bioinformatics*. 9:384.

Royle L, Campbell MP, Radcliffe CM, White DM, Harvey DJ, Abrahams JL, Kim YG, Henry GW, Shadick NA, Weinblatt ME, et al. 2008. HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. *Anal Biochem*. 376:1–12.

Rudd PM, Elliott T, Cresswell P, Wilson IA, Dwek RA. 2001. Glycosylation and the immune system. *Science*. 291:2370–2376.

Saldova R, Royle L, Radcliffe CM, Abd Hamid UM, Evans R, Arnold JN, Banks RE, Hutson R, Harvey DJ, Antrobus R, et al. 2007. Ovarian cancer is associated with changes in glycosylation in both acute-phase proteins and IgG. *Glycobiology*. 17:1344–1356.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 37:D5–15.

Stenutz R. 2009. Automatic spectrum interpretation based on increment rules. In von der Lieth CW, Lütteke T, et al., editors. *Bioinformatics for Glycobiology and Glycomics: An Introduction*. Wiley-Blackwell. p. 311–320.

Taniguchi N, Hancock W, Lubman DM, Rudd PM. 2009. The second golden age of glycomics: From functional glycomics to clinical applications. *J Proteome Res*. 8:425–426.

Tharmalingam T, Marino K, Rudd PM. 2010. Platform technology to identify potential disease markers and establish heritability and environmental determinants of the human serum N-glycome. *Carbohydr Res*. 345:1280–1282.

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*.

Toukach FV, Knirel YA. 2005. New database of bacterial carbohydrate structures. Proceeding of the XVIII International Symposium on Glycoconjugates, 216–217.

Vanderschaeghe D, Festjens N, Delanghe J, Callewaert N. 2010. Glycome profiling using modern glycomics technology: Technical aspects and applications. *Biol Chem*. 391:149–161.

Velankar S, Best C, Beuth B, Boutselakis CH, Cobley N, Sousa Da Silva AW, Dimitropoulos D, Golovin A, Hirshberg M, John M, et al. 2010. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res*. 38:D308–D317.

von der Lieth CW, Bohne-Lang A, Lohmann KK, Frank M. 2004. Bioinformatics for glycomics: Status, methods, requirements and perspectives. *Brief Bioinform*. 5:164–178.

von der Lieth CW, Lütteke 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.

von Witzendorff D, Maass K, Pich A, Ebeling S, Kolle S, Kochel C, Ekhlasi-Hundrieser M, Geyer H, Geyer R, Topfer-Petersen E. 2009. Characterization of the acidic N-linked glycans of the zona pellucida of prepuberal pigs by a mass spectrometric approach. *Carbohydr Res*. 344:1541–1549.

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*. 59:687–696.

Widmalm G. 2007. General NMR spectroscopy of carbohydrates and conformational analysis in solution. In Kamerling JP, editor. *Comprehensive Glycoscience*. Oxford: Elsevier, p. 101–132.

Zaia J. 2008. Mass spectrometry and the emerging field of glycomics. *Chem Biol*. 15:881–892.