Supporting Information
for Proteomics
DOI 10.1002/pmic.200700917

Nicolle H. Packer, Claus-Wilhelm von der Lieth, Kiyoko F. Aoki-Kinoshita,
Carlito B. Lebrilla, James C. Paulson, Rahul Raman, Pauline Rudd,
Ram Sasisekharan, Naoyuki Taniguchi and William S. York

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)
Supporting Information

1. GLYDE-II: Exchange format for glycan structures

It is clear that the amount and diversity of glycomics data make it necessary to distribute it in different locations throughout the world. This approach allows those having the technical expertise required for data generation to maintain close ties with the data and its curation. Furthermore, a comprehensive collection of data processing tools must be available to interpret and mine the data. Maintenance of this tool collection at a single site would be difficult, but these tools can also be distributed over multiple sites that are maintained by the tool developers themselves. This distributed approach makes it necessary to transmit data over the Internet when it is being used for biomarker discovery or other purposes. Such a system can work only if robust standards for data transmission are developed and accepted by the scientific community. A primary requirement for biomarker discovery that was recognized by the workshop participants is the development of a standard format for the exchange of glycan structural data over the Internet.

Web services (WSs) are an emerging technology that can enable such a distributed system for data archiving, retrieval, and processing. WSs are “software systems designed to support interoperable machine-to-machine interaction over a network.” (See http://en.wikipedia.org/wiki/Web_services). Extensible Markup Language (XML—http://www.w3schools.com/xml/default.asp) is the de facto standard for the exchange of data by Web services, as it is specifically designed to represent data in a way that can be easily parsed. Therefore, XML is the natural choice for developing a standard for the exchange of carbohydrate structural data over the Internet.

The exchange of XML-encoded structural data will enable Web services or isolated software applications to process the data, thus making it possible to evaluate complex, high-level queries generated by users of the system. In the near future, it will be possible to evaluate such queries using Web processes, which are workflows that are implemented in a distributed fashion over the Internet. Technology for the automatic discovery of Web services and their incorporation into Web processes is a very active area of research in the computer science community, and significant progress is being made in this area. Virtually all of this research involves the use of XML as a common data exchange language.

Biomarker discovery requires analysis of highly complex samples of tissues or body fluids. The complexity of these samples is due, in part, to the fact that many different glycan species are present, and each of these can be attached to many different proteins, peptides, lipids, or small molecules. Since the number of resulting glycoconjugates can be astronomical, a modular representation of glycans, lipids, proteins, and small molecules as separate entities and glycoconjugates as a combination of these modules is required. At one level, the structure of a glycoconjugate can be indicated by simple pointers to its individual component modules in the databases. For example, the structure of a specific glycoprotein can be succinctly represented as a pointer to the glycan in a glycan structure database, a pointer to a protein in a protein database, and a description of the amino acid to which the glycan is attached. However, many analyses will require an explicit description of the molecular structure of the glycoconjugate. Imagine a protein database that contains only accession numbers (pointers) but does not provide a mechanism to represent and exchange data describing the amino acid sequence of the proteins. This would not be very useful, as routine tasks like BLAST searches would be impossible without an explicit representation of the sequence. The workshop participants agreed that a robust data exchange format that can explicitly represent glycoconjugate structural information at several levels of granularity is highly desirable.


Of these, only the last three are written in XML. CML is an atomistic representation based on a connection table (also called an adjacency list, http://en.wikipedia.org/wiki/Adjacency_list), but does not provide for the abstraction of monosaccharide residues, which are used as the basic building blocks of complex glycans in nearly all other formats. Conversely, CabosML and GLYDE are based on abstracted monosaccharide residues. These formats are based on a tree-like formalism (http://en.wikipedia.org/wiki/Tree_%28data_structure%29) rather than a connection table.

The GLYDE version 1 format has been available for evaluation for approximately 1 year, and during that time different research groups have noted that the tree formalism employed by both GLYDE and CabosML imposes restrictions that hinder the specification of some glycan structures and also places limits on the use of description logics to semantically describe the monosaccharide residues. Therefore, the consensus of GLYDE users was that a connection table
approach is more appropriate for a structural data exchange format.

This led to the initial development of GLYDE-II, an XML representation of the chemical structures of biological molecules (including macromolecules) that is based on a connection table (CT) formalism. The goal of GLYDE-II is to provide a mechanism to completely and unambiguously specify the complete structure of biological molecules (including complex glycans) at several levels of granularity.

The participants of the workshop agreed that GLYDE-II should be adopted as the standard for glycan structural data exchange to be used by the distributed glycomics analysis software and databases. Scientists at the German Cancer Research Center (DKFZ), the Consortium for Functional Glycomics (CFG), and the Complex Carbohydrate Research Center (CCRC), who are coauthors of this white paper, have agreed to collaborate to develop a fully functional version of GLYDE-II for use by the glycomics community in the near future.

2. Background information to support the decisions

Here we present additional information regarding the status of the currently existing databases and the background for priority two and three.

2.1 Current status of carbohydrates containing databases

With 49,897 entries, which correspond to 23,118 distinct 2-D graphs, the CCSD (CarbBank) is still the largest publicly available repository of glycan-related data. All consecutive open access projects, which were started in the USA (CFG), Europe (GLYCOSCIENCES.de), and Japan (KEGG-Glycan) at the beginning of the new century, made use of the CCSD data and included a subset of all CCSD entries in their newly designed databases. However, none of these new initiatives aimed to provide the scientific community with a service similar to that provided by CarbBank by providing access to all published carbohydrate structures. CFG, KEGG, and GLYCOSCIENCES added only a limited amount of new structures to complete the newly established databases according their specific focus. An exception is the Russian Bacterial Carbohydrate Structure Database (BCSDB) which follows the CarbBank mission and claims to cover nearly all structures of “bacterial carbohydrates” published before 2006 (currently 5904 structures, 2310 are taken from CCSD). The Australian company Proteome Systems developed the commercially available GlycoSuiteDB, which followed the CarbBank data model; however, GlycoSuiteDB seems to provide a more rigorous handling of the associated biological annotations. Release 8.0 (August 2005) contains 9436 entries, sourced from 864 references with 3238 unique glycan structures, of which 1851 are completely characterized. However, there is not yet open access to their data.

A recent analysis of all publicly available glycan structures (CFG, KEGG-Glycan, BCSDB, and GLYCOSCIENCES.de) revealed that all newer projects have (i) converted only a subset of all CarbBank structures, (ii) have not completely transferred all associated data provided as free text, (iii) made mistakes when converting the CarbBank 2-D structure graph into the Internal data format, which is unique for each project (Linear Code, KCF, IUPAC adopted representation, LINUCS). We estimate that the error rate is in the range of 5–10% of all converted entries.

Additionally, the original CarbBank data contain various inconsistencies: the linkage patterns do not follow a strict pattern, monosaccharide names contain various types of errors, and the free text fields need additional curation to derive a controlled vocabulary. We estimate that about 80% of all CarbBank entries can be reliably converted using automatic procedures.

2.2 Support for distributed, federated databases providing tools to store the experimental data that support the identification of specific structures in biological samples

Since the Internet offers a unique chance to constitute a global and interactive peer-to-peer network for the exchange for scientific data, open access networks will provide an effective platform to install federated databases. The EUROCarbDB project, a design study funded by the EU, is currently attempting to work out this model. It aims to create the foundations for databases and bioinformatics tools in the realm of glycobiology and glycomics, and will establish mainly the technical framework for a bottom to top initiative where all interested research groups can feed in their primary data. The new infrastructure will constitute the nucleus for the creation of a depository for carbohydrate-related data similar to the extensively used data collections in the area of genomics and proteomics. The EUROCarbDB design study concentrates on the evaluation and development of the basic requirements for the proposed infrastructure.

(i) The predefinition of standard representations of both the methods used and the data generated in glycobiology/ glycomics studies, guidelines for good practice, establishment of procedures for quality control, and the development of associated data base models.

(ii) The design of software tools, which enable a peer-to-peer network of federated databases for glycomics. The availability of such a tool will encourage people to input their recorded experimental data into a local database that may be kept private until it is published. Inexpensive hardware platforms and the availability of free software tools will favor this process.
(iii) The development of algorithms that enable the rapid and reliable automatic interpretation of MS and HPLC data and NMR spectra. The existence of sufficiently large high-quality collections of reference spectra is an urgent demand of ongoing high-throughput glycomics projects.

However, a central depository for high-quality primary experimental reference data might be desirable and should also be considered. The installation of a central infrastructure is probably the most trustable solution to guarantee the integrity of reference data as well as their long-term availability.

It was agreed that significant progress in the area of structural databases and internal representations at the EuroCarbDB make this a logical place to house the unified structural database. Development of this resource will facilitate integration of diverse data collections (such as mass spectral data) housed at different institutions.

2.3 Support for the development of open source software projects for glycomics

As discussed above, the current state of glyco-related databases can be characterized as "the biggest defect in the field" (Ajit Varki). Therefore, highest priority has been assigned to the database gap. However, it is obvious that efforts to form robustly interactive databases must be combined with efforts to develop effective user-friendly software to access and analyze the data deposited in the emerging databases.

Currently, only a limited amount of freely available software related to glycomics is available to be shared by various projects. The largest bottleneck has been the lack of a common language to exchange glycan structures and related data. With the agreement to accept GLYDE-II as the central format for exchanging structural data, a central prerequisite for the development of glyco-related software engineering resource glycomics has been achieved.

Experience in other fields of software development and engineering is that the Open Source philosophy favors the rapid creation of robust solutions within an open, collaborative environment. Its underlying philosophy is that source code is available for anyone to use, modify, and redistribute freely. Open Source projects have been shown to promote a higher standard of quality, and help to ensure the long-term viability of both data and applications. The collaborative spirit of all the major glycomics initiatives and their willingness to collaborate and integrate their individual resources as expressed during the ‘Frontiers in Glycomics’ meeting provide an excellent starting point for the success of an Open Source initiative in glycomics.

2.3.1 The areas of software development where solutions are urgently required

(i) Implementation of the GLYDE-II format. The major currently existing formats should be supported.

(ii) User-friendly graphical interfaces (GUIs) that allow manual input of glycan structures and controlled semantic annotation of related data. The availability of such tools is also a paramount requirement to convince scientists/editors/publishers that structural and primary experimental data should deposited in DBs as part of the publication process.

(iii) Robust graphical interfaces to input and retrieve all types of glycan structures in the appropriate biological context.

(iv) Tools for reading, processing, and annotation of experimental data (e.g., MS, HPLC, and NMR data origination from various spectrometers).

(v) Algorithms for robust automated annotation and interpretation of experimental data, especially data generated by key analytical techniques (MS, HPLC, and NMR).

(vi) Quality scoring schemata for all types of experimental and theoretical data. These may be different for submission to journal or to a DB.

2.3.2 Medium-term project of high interest

(i) Creation of robust tools and services for automated annotation and interpretation of experimental data, including MS and NMR spectra as well as HPLC profiles. Similar to the well-established and routinely used ‘Protein mass fingerprinting’ services in proteomics (e.g., MASCOT or SEQUEST) the establishment of corresponding freely accessible ‘glycan mass fingerprinting’ services are highly desirable.

(ii) A portal site for access to multiple carbohydrate-related data such that a single query may be sent to all relevant databases for retrieval.

(iii) Tools for data mining, e.g., substructure and similarity searches for annotating glycan (sub)structures; correlation of genomic data with glycan data (e.g., gene expression of glycosyltransferases and MS data of identified glycans); statistical analysis of glycan structures found in various species; mining glycan and related data in an integrated manner using advanced approaches such as kernel methods and probabilistic models; and analysis of mass amounts protein–carbohydrate interaction data, e.g., binding affinity using frontal affinity chromatography and glycan arrays.

(iv) Development of robust ontologies for glycomics, which will favor the development of: automatic procedures for semantic annotation of high-throughput data; access to controlled vocabulary; dynamic connection to other disciplines; and improved data mining approaches.

(v) Development of simulation approaches: to explore the structural variability of glycans produced by various species; to explore the conformational space glycans in its physiological surroundings; and to find potential binding partners of glycans and identify the spatial location where the glycans bind.
The outlined tasks are huge and definitively cannot be accomplished by the few research groups currently working in this emerging area. It is of paramount importance to attract additional bioinformatics scientists by providing freely available, well-structured, and scientifically relevant data along with open software tools for manipulating that data.

2.3.3 Other important tasks, which can be best organized through a collective effort by the glycosciences community

A larger collaborative endeavor that requires input of the scientific community is the establishment of shared definitions and guidelines for quality/reliability metrics for evaluation of various types of experimental data. It is necessary to raise awareness within the community that the standardization of structural descriptions, establishment of common experimental protocols, and the implementation of appropriate quality scoring schemata are urgent tasks that are necessary to advance the field of glycomics/glycobiology. This is especially true for emerging high-throughput procedures.

Another area where the glycosciences community can be harnessed is the curation and revision of the databases in response to advances in the field. This can be fostered by a community spirit that encourages and honors well-recognized experts, younger investigators, and retired but still active members of the glycobiology community who actively contribute to the maintenance of the integrity and timeliness of these databases.

Modern Internet techniques massively favor such a worldwide distribution of tasks to collect knowledge and maintain the consistency of data. One way to foster the use of the Internet as a resource that unites glycoscientists is the development of a Web-based encyclopedia for glycobiology/glycomics that is implemented using the Wikipedia model.