Advancing glycomics: Implementation strategies at the Consortium for Functional Glycomics

Rahul Raman, Maha Venkataraman, Subu Ramakrishnan, Wei Lang, S. Raguram, and Ram Sasisekharan

Bioinformatic platforms to acquire, integrate, and disseminate these data sets in a meaningful fashion. With the consortium for functional glycomics (CFG) as the model system, this review discusses databases and the bioinformatics platform developed by this consortium to advance glycomics.

Key words: bioinformatics/CFG/glycan/glycomics

Introduction

In comparison with genomics and proteomics, advancing glycomics is faced with unique and intriguing challenges. These challenges arise from two fundamental aspects of glycan structure–function relationships. First, the biosynthesis of glycans is a non-template-driven process involving a coordinated expression of multiple glycosyltransferases, some of which have additional tissue-specific isoforms (Varki et al., 1999; Lowe and Marth, 2003; Taylor and Drickamer, 2003). Second, understanding the biochemical basis of glycan–protein interactions in the context of a biological pathway is complicated by the multivalency and the graded affinity involving an ensemble of glycan structures making multiple contacts with multivalent binding sites on proteins (Collins and Paulson, 2004). The above issues have also made it challenging to develop databases and bioinformatics tools for glycomics (von der Lieth et al., 2004; Raman et al., 2005).

Based on the above, the evolution of information in glycomics has been different from that of genomics and proteomics. Although large volumes of DNA and protein sequence data were being generated in databases such as GenBank and SwissProt, much less information on glycan structures was available at that time. For example, early attempts to construct a glycan database resulted in the Complex Carbohydrate Structure Database (CCSD) which contained about 2000 literature citations (Doubet et al., 1989) along with a set of computational tools (CarbBank) to query and access information. CCSD rapidly grew into the primary repository for glycan structures with over 50,000 citations. However, challenges in structural characterization of glycans in the past led to numerous glycan structures with unspecified linkage information which complicated the databasing and search engine development. Moreover, during the development of CCSD, a clear blueprint was not available for establishing the kinds of data that needed to be stored and annotated along with the glycan structures in this database. Owing to these and other challenges, further development of CCSD was discontinued.

Important breakthroughs in glycobiology fueled by rapid technology development have led to a “glyco-renaissance” in the past few years. Recognizing the need to take an integrated approach to understand glycan structure–function relationships, several international collaborative efforts such as the Consortium for Functional Glycomics (CFG; an international initiative funded by National Institute of General Medical Sciences), Complex Carbohydrate Research Center (www.cccr.uga.edu), EuroCarbDB (www.eurocarb.org), and Human Disease Glycomics/Proteome Initiative (www.hpgi.jp) have been established. Motivated by the need to address the challenges in glycomics, these collaborative efforts are developing novel resources and state-of-the-art technologies for advancing this field. Importantly, these initiatives are putting significant resources into developing databases and bioinformatics platforms to integrate and disseminate glycomics data to the scientific community. This review discusses the utility of the glycomics databases developed by the CFG highlighting the CFG’s synergistic approach to collaborate with other large-scale initiatives and collectively address challenges in advancing glycomics.

CFG bioinformatics platform for glycomics

To address the central issue of decoding structure–function relationships of glycan–protein interactions, the CFG is organized into scientific Cores that have developed technologies to generate novel data sets. These diverse data sets are derived from (1) gene expression of glycosyltransferases and...
glycan-binding proteins (GBPs), (2) phenotyping analysis of transgenic mice, (3) mass spectrometric profiling of glycan structures isolated from cells and tissues, and (4) screening glycan affinity of proteins using novel glycan arrays. It is clear that there is a need to cut across these diverse data sets to begin understanding the fundamental structure-function relationships of glycans. A critical component that enables this process is a bioinformatics platform to store, integrate, and process the information generated by the above methods and disseminate them in a meaningful fashion via the Internet to the scientific community worldwide.

To implement an informatics framework for integrating diverse data sets, the Bioinformatics Core (Core B) of the CFG constructed relational databases. The blueprint of the CFG database is the data model or ontology diagram that captures data definitions and inter-relationships (Figure 1B). It is important to keep the complexity of this database not come in the way of the researchers who deposit data into the database and query the database for accessing specific data sets. To accomplish this goal, the CFG bioinformatics platform has developed a three-tier architecture. This comprises of the relational database in the backend (implemented using

---

**Fig. 1.** CFG bioinformatics platform. Shown in panel A is the three-tier architecture developed to provide flexibility to handle evolving data relationships and to facilitate seamless data acquisition and dissemination. Shown in panel B is a schematic of the data model where the key data objects and their connectivity with other objects are indicated.
Scientific core data dissemination

As stated above, there are four primary data sets generated by the CFG scientific Cores. Each of these data sets comprise of hierarchical layers of information starting from a summary presentation (in PDF or Excel format) of the data to the individual parameters and the data files. The dynamic data dissemination interfaces facilitate the navigation of data sets through these hierarchical layers giving users access all the way down to the individual raw data files. Furthermore, the standardized protocols for performing the different analyses by each scientific Core are also available on the front page of each of the dissemination interfaces.

Glycan analysis data

The Analytical Glycotechnology Core utilizes the matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) methodology to obtain profiles of glycans derived from tissues and cells of mice (wild-type [Comelli et al., 2006] and knock-out strains) and humans. Each annotated glycan structure in the MALDI-MS profile represents the best level of information captured based on its mass, composition, biosynthesis, and biological source. More detailed analysis such as tandem MS-MS fragmentation of specific mass ions (these ions are highlighted in a shaded box in the MALDI-MS profile) is used to resolve isobaric structures. The MALDI-MS profiling data provides a high throughput snapshot of the most likely glycan structures derived from specific cells and tissues.

The current dissemination interface of the glycan analysis data organizes the sample information at different hierarchical levels from species→tissue→sample→N-/O-linked glycan profile→high/low molecular weight glycans (in cases where high and low molecular weight glycans are analyzed separately). The final annotated spectrum is available in both JPG and PDF for viewing, comparison, and printing purposes. In addition to the annotated spectra, the ASCII file containing m/z and intensity values and the raw binary format which is generated by the MALDI-MS instrumentation is also accessible to the user. Currently, the annotated image files are being converted to a digitized format comprising of the m/z peak, intensity, IUPAC representation of the most likely set of glycan structures, and links to corresponding structures in the glycan structures database. The novel structures that are obtained by rigorous structural characterization methods such as tandem MS² and gas chromatography-MS linkage analysis are entered directly into the glycan structures database.

Glyco-gene microarray data

The CFG has developed glyco-gene DNA microarrays to analyze the expression of genes for GBP and glycan biosynthesis enzymes. The CFG glyco-gene microarrays have been successfully utilized by many participating investigators (Smith et al., 2005; Comelli et al., 2006) to analyze a wide range of samples such as different tumor cell lines, cells from wild-type and knock-out mice strains, cells under mechanical stress, and cells in other pathological states.

The data obtained from the glyco-gene DNA microarray analysis is organized based on the scope of the experiment, information on the samples used in the experiment (in the MAIME standard format), and the various data files. The data files include the standard set of Affymetrix files—.CEL, .CHP, .DAT, .RPT and .EXP—and an ASCII-based read-out file containing probeset ID, signal intensity, present/absent call, and expectation value for each probe in the gene chip. The ASCII-based readout file is translated by the middleware application layer, and the information regarding the

---

Table I. URLs for the CFG data dissemination pages

<table>
<thead>
<tr>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFG home page</td>
<td><a href="http://www.functionalglycomics.org">http://www.functionalglycomics.org</a></td>
</tr>
<tr>
<td>GBP molecule pages</td>
<td><a href="http://www.functionalglycomics.org/glycomics/molecule/jsp/gbpMolecule-home.jsp">http://www.functionalglycomics.org/glycomics/molecule/jsp/gbpMolecule-home.jsp</a></td>
</tr>
<tr>
<td>Glycan database</td>
<td><a href="http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp">http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp</a></td>
</tr>
<tr>
<td>Glycosylation pathways</td>
<td><a href="http://www.functionalglycomics.org/static/gt/gtdbs.shtml">http://www.functionalglycomics.org/static/gt/gtdbs.shtml</a></td>
</tr>
<tr>
<td>Glycan profiling data</td>
<td><a href="http://www.functionalglycomics.org/glycomics/publicdata/glycoprofiling.jsp">http://www.functionalglycomics.org/glycomics/publicdata/glycoprofiling.jsp</a></td>
</tr>
<tr>
<td>Glycan array screening data</td>
<td><a href="http://www.functionalglycomics.org/glycomics/publicdata/primaryscreen.jsp">http://www.functionalglycomics.org/glycomics/publicdata/primaryscreen.jsp</a></td>
</tr>
<tr>
<td>Gene microarray data</td>
<td><a href="http://www.functionalglycomics.org/glycomics/publicdata/microarray.jsp">http://www.functionalglycomics.org/glycomics/publicdata/microarray.jsp</a></td>
</tr>
<tr>
<td>Mouse line phenotyping data</td>
<td><a href="http://www.functionalglycomics.org/glycomics/publicdata/phenotyping.jsp">http://www.functionalglycomics.org/glycomics/publicdata/phenotyping.jsp</a></td>
</tr>
</tbody>
</table>
expression of each gene is automatically captured in the tables of the relational database. The dissemination interface provides a structured navigation of the microarray data sets starting from the experiment and sample information to downloading individual data files for each sample. Since the expression data is captured at the level of each gene represented on the microarray, tools to download the expression for a specific set of genes across the different samples for a given experiment have also been implemented as a part of the dissemination interface. Furthermore, the dissemination interface also includes links to perform low and high level data analysis using the downloaded microarray data. The user-friendly dissemination of the gene microarray data has motivated efforts to develop data-mining tools for the prediction of glycan structures based on gene expression profile (Kawano et al., 2005).

**Mouse phenotype data**

Advances in whole organism genetics have provided insights into linking the role of glycosylation and glycan diversity to whole organism phenotype (Lowe and Marth, 2003). Motivated by these efforts, the CFG dedicated two scientific Core facilities to generate transgenic mice with knockouts in glycosyltransferase and GBP genes (Mouse Transgenics Core) and to perform a battery of phenotyping analyses on these transgenic mice (Mouse Phenotyping Core). The analyses done by the Mouse Phenotyping Core can be classified into hematology, histology, immunology, metabolism, and behavior, and they provide a large volume of new data at the whole organism level. The data sets include fluorescence activated cell sorting analysis, histological staining of different tissues, and a wide range of parameters pertaining to oxygen consumption, motor reflexes, blood count, and coagulation. As a result, diverse data formats such as PDF, JPG/TIFF images, and Excel spreadsheets are uploaded into the database and are automatically organized into the appropriate tables. In addition, a summary on each of the broad type of analysis, which contains specific interpretations of the Core Director along with the most significant graphs or parameter values, is also accessible via the database.

Similar to the other Core dissemination interfaces, the topmost layer of the mouse phenotyping data interface comprises of a list of transgenic mice that have been phenotyped along with the links to data summaries, experimental details, and raw data files. The experimental details provide a description of the tests performed and comprehensive information on the mice used in the studies. The raw data provide access to individual parameters and image files (e.g., histological staining of individual tissues).

**Glycan–protein interaction screening data**

To expand the current knowledge of sequence-specific glycan–protein interactions, the Glycan Synthesis and Glycan–Protein Interaction Cores of the CFG are developing glycan arrays comprising of hundreds of synthetic and biological glycans. Two different types of array formats have been developed. The first format is a microwell-based array where the glycans are biotinylated and applied to streptavidin-coated microwells. The second format is a N-hydroxysuccinimide activated glass slide array where the glycans are covalently printed on the slide using standard printing technologies available for making DNA microarrays (Blixt et al., 2004). The printed array has better signal to noise ratio and also facilitates expansion of the number of glycans that can be printed on the slide. These glycan arrays are becoming widely utilized by the scientific community for screening several proteins such as plant and animal lectins, antibodies, and proteins on pathogen cell surfaces to identify novel glycan ligand specificity of these proteins (Guo et al., 2004; Bochner et al., 2005; Tateno et al., 2005; van Vliet et al., 2005; Singh et al., 2006). More recently, a new technique to derivatize the glycans using a fluorescent tag has enabled developing glycan arrays with the quantification of each glycan on the array (Xia et al., 2005).

Annotation tools have been developed to automatically parse the readout files derived from the screening analysis and store the information on the glycan structure, mean signal intensity, signal to noise ratio, and standard error in the database. At the top level of the data dissemination page, the information is organized into the protein analyzed, experiment information, and the name of the investigator. The data prepared by the Core are available as an Excel format with some of the legacy data containing PDF data summaries. An interactive interface that provides a two-dimensional false color imaging of the array (based on signal intensities) and a bar chart of signal intensity versus glycan ID is also available. This interface allows the user to seamlessly navigate from the click on potential high signal intensity regions (in the two-dimensional representation and bar graph) to provide more information on that glycan structure in the glycan structures database.

In addition to the hierarchical dissemination of the diverse data sets generated by each scientific Core, the relationships between data sets at the level of the sample are also captured in the relational database (Figure 2). For example, the CFG has begun integrating its tools to derive orthogonal data sets such as gene expression profile and glycan analysis on tissues and pure cell populations derived from human and mouse (Comelli et al., 2006). Such integration enables researchers to begin correlating glycan diversity of a cell or tissue with the gene expression profile of the glycan biosynthesis enzymes and also to understand the glycan ‘signature’ of phenotypically distinct cells and tissues derived from different mice strains.

**GBP molecule pages**

An emerging concept in data integration is the molecule page interface which provides a portal to information and data ranging from molecule to mouse (Li et al., 2002; Raman et al., 2005). The molecule page interface developed by CFG captures information pertaining to different families of human and mouse GBPs. The CFG classification of GBP families is—C-type lectins, galectins, siglecs, and other. The CFG molecule pages contain three main components (1) automatic acquisition of information from other public databases on that molecule, (2) automatic interface with
CFG data pertaining to that molecule, and (3) contribution from experts on that particular molecule. The GBP molecule pages are organized based on the above classification and subfamilies. Each class is linked to a list of proteins in that class, and each protein is linked to its appropriate molecule page. The information pertaining to the protein is organized into six categories that are presented as clickable tabs in the molecule page interface.
The General tab comprises of name of the molecule, its synonyms obtained from the SwissProt database, and a summary of the molecule that is contributed by experts (work is in progress to obtain expert contribution for molecule pages). The Reference tab comprises of link to NCBI’s PubMed database that automatically searches this database using the name of the protein and synonyms as search fields. It also contains links to other portals such as NCBI’s Entrez Gene, Source data portal (developed at Stanford University). The Genome tab has information on the gene name, the cDNA sequence with links to BLAST server for sequence alignment, gene expression profiles available in public databases such as NCBI’s Gene Expression Omnibus database, and Symatlas database of Novartis foundation. Currently, work is in progress to interface the gene expression data obtained from the CFG’s glyco-gene microarrays.

The Proteome tab has information pertaining to the protein sequence with links to the SwissProt and the PDB databases (if three-dimensional crystal structures are available). Furthermore, a schematic of the domain organization of the entire protein family is provided. The Glycome tab is an unique contribution from the CFG which provides information on candidate ligands or known counter receptors for that GBP (provided by expert contribution) and access to glycan array data if that protein was screened on the CFG glycan array (Figure 3). More recently, using the tools developed to extract glycan structures from PDB (Lutteke et al., 2004), information on the glycan ligands used in the crystal structure studies of that GBP (if available) with links to the appropriate PDB entries is provided. Finally, the Biology tab provides a summary of physiological and pathological roles of that GBP and an interface to CFG phenotyping analysis (if available) of transgenic mice with a knockout of that protein.

Glycan structures database

The CFG glycan structures database represents one of the many important efforts (Cooper et al., 2001; Hashimoto et al., 2005; Lutteke et al., 2005) to develop a standardized repository of glycan structures information. This database was developed to meet three main objectives. The first objective was to facilitate the assignment of peaks in the MALDI-MS glycan profiling of tissues. The second objective was to capture relationship between candidate ligands on the glycan array and their corresponding binding proteins. The third objective was to capture information on glycan structures that are being published in the literature (Figure 1B). The current repository was built starting from mammalian structures in the CCSD, curated structures obtained from a private database (developed by Glycominds Ltd., Lod, Israel). To this repository, the glycan structures generated by CFG (Glycan Synthesis Core and those on the glycan array) were added along with synthesis protocols for each glycan synthesized by Glycan Synthesis Core. This database also comprises of a large number of theoretically generated mammalian N-linked glycans based on biosynthesis rules that are primarily utilized for annotation of the MALDI-MS glycan profiles.

As stated earlier, the glycan array data have already been integrated with the glycan structures database. Upon accessing a glycan on the glycan array, the list of proteins for which that structure was identified as a high-affinity binder is also available along with the entire data set of that screening experiment. This integration not only enables the identification of a high-affinity glycan ligand for a given protein, but it also provides information on what other proteins were identified as high-affinity binders to the same glycan ligand. Currently, efforts are ongoing to integrate the glycan-profiling data with the glycan structures database.

The glycan structures in the database can be searched and retrieved using different search criteria such as molecular weight, composition, biological source, linear nomenclature, and citation. Another useful feature is the substructure search interface that allows users to either build from common templates (core and extension) or modify imported glycan substructures or motifs and search for these motifs in the database. This interface also facilitates development of glycan biosynthesis pathway interface to the glycan structures database (see Glycosylation Pathways Interface) and entering a new glycan structure into the database.

Given that the glycan structures databases represent a common focus of all of the large glycomics initiatives, there is a practical need to develop standardized formats for exchange of information on the structures. Collaborative efforts between large-scale initiatives are in progress to evaluate XML formats (Kikuchi et al., 2005; Sahoo et al., 2005) for consistent description of glycan structures in different databases.

Glycosylation pathways interface

Understanding the glycosylation pathways involved in glycan biosynthesis is the primary link to understand the structure–function relationship of glycans in the context of how genotype influences whole organism phenotype. The KEGG (Hashimoto et al., 2005) and the CAZy (http://afmb.cnrs-mrs.fr/CAZY/) databases are currently the major sources of information on glycan biosynthesis enzymes. The CFG glycosylation pathways interface provides a set of composite glycan structures representing the core, terminal, and common extension units of N-linked, O-linked glycans and glycolipids. Each linkage in the composite structure interface is directly linked to the corresponding family of the glycosyltransferase in the CAZY database. The KEGG database provides a pathway representation for the biosynthesis of different glycans. There are 98 genes corresponding to human glycosylation pathways that have been annotated (assigned to a specific glycosidic linkage) in the KEGG database.

To expand this annotation, the CFG is collaborating with CAZy, KEGG, and other experts. The current list comprises around 200 annotated genes which will soon become available via the KEGG and CFG databases. To facilitate the inputs from the different initiatives, experts in the CFG have constructed detailed composite structures of core, extension, and terminal regions of different glycans. Each monosaccharide in this composite structure and the linkage between the specific monosaccharides are numbered.

Advancing glycomics
Fig. 3. GBP molecule pages. Shown is the example of the Glycome tab of the molecule page of human dendritic cell intercellular adhesion molecule grabbing nonintegrin (DC-SIGN) which is a type II C-type lectin. The information on primary glycan specificity and proposed glycoprotein/glycolipid counter receptors is provided by expert investigators. Also shown is the link to the glycan array data generated for this protein which is automatically associated with the molecule page. Finally, the glycan ligands extracted from the PDB with the PDB identifiers of the protein–glycan complexes are shown.
Advancing glycomics

Summary

The beginning of this millennium has marked a promising era for advancing glycomics where large-scale international initiatives such as CFG are generating valuable resources and data sets that are openly available to the scientific community. The establishment of these initiatives has sparked an explosive growth of novel contributions to the glycomics field. Central to the progress of these efforts is the development of bioinformatics platforms to acquire and disseminate diverse data sets across the world using user-friendly interfaces via the Internet. Another important aspect of these large glycomics initiatives is the motivation to take a collaborative approach for integrating the resources and data generated by the different initiatives as pointed out by the Editorial (2005) in Nature Methods. This collaboration and integration is critical for providing access to diverse data sets in different databases that enable scientists to begin answering important questions on glycan diversity, glycan–protein interactions that are fundamental to understanding the structure–function relationships of glycans.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences Glue Grant U54 GM62116. The authors thank present and past members of the CFG Bioinformatics Core including Ganesh Venkataraman, Chipong Kwan, Eric Berry, Nishla Keiser, and Ishan Capila. In addition, the authors thank the members of the other Cores, the Steering Committee, and the participating investigators of the CFG for their contribution to the Bioinformatics Core.

Conflict of interest statement

None declared.

Abbreviations

CCSD, Complex Carbohydrate Structure Database; CFG, consortium for functional glycomics; GBP, glycan-binding

Glycan Nomenclature

- Mannose
- Galactose
- N-acetyl glucosamine
- N-acetyl galactosamine
- Glucuronic acid
- N-acetyl neuraminic acid
- Fucose
- 3S 3-O sulfate
- 6S 6-O sulfate

Fig. 4. Composite structure glycosylation pathways. The composite structures of complex type N-linked glycan along with the common type II extension unit and the known terminal monosaccharides are shown. Each monosaccharide is numbered to facilitate the annotation of the family of glycosyltransferases that are involved in the synthesis of the linkage between specific monosaccharides. The glycosyltransferase molecule pages are being developed and would be accessible by clicking on specific linkages in the composite structure interface.
protein; MALDI-MS, matrix-assisted laser desorption/ionization mass spectrometry.

References


