An integrated systems approach to structure-function relationships of glycans[11]

Final Paper

for

Genomics and Bioinformatics Fall 2005

Instructor	**: Prof. Mark Gerstein**
TA	**: Sara Nichols & Xiaowei Zhu**
Student	: Hugo. Y. K. Lam
Student ID	: 904907866
Date	: December 10, 2005

Contents

SECTION ONE: REVIEW 3

INTRODUCTION 3
CONCEPTS AND METHODS 3
FUNCTIONAL GENETICS APPROACH TO GLYCOMICS 3
DEVELOPMENT OF GLYCOL-GENE MICROARRAY FOR GLYCOMICS 4
GLYCAN ANALYSIS – FROM HIGH-THROUGHPUT TO FINE STRUCTURE CHARACTERIZATION 4
BIOCHEMICAL ANALYSIS OF SPECIFICITY OF GLYCAN-PROTEIN INTERACTIONS 4
BIOINFORMATICS PLATFORM FOR GLYCOMICS 5
COMMENTS 6

SECTION TWO: PROPOSAL 8

INTRODUCTION 8
INTEGRATION APPROACHES 8
APPLICATION SPECIFIC SOLUTIONS 8
WORKFLOW SYSTEMS 9
DATA WAREHOUSING AND DATA FEDERATION 9
SEMANTIC WEB 9
A SEMANTIC WEB PLATFORM 10
METHODOLOGY 10
METHODS 12

REFERENCES 13

Section One: Review

Introduction

Glycosylation is the process or result of adding one or more sugars to a protein and lipid[1]. It is perhaps the most extensive protein post-translation modification which generates extensive phenotypes from a limited genotype, for providing the essential functional diversity to mammals. Glycans are found ubiquitously present at the cell-extracellular interface to modulate protein activity.

There are three fundamental and interrelated aspects which complicate the study of glycans. First, the biosynthesis of glycans is a non-template driven process which leads to inherent heterogeneity and large diversity of glycan structures. Second, this property has challenged the development of analytical techniques to accurately define their chemical structures. Finally, the multivalency and graded affinity involving an ensemble of glycans making multiple contacts with multivalent protein binding sites has complicated the understanding of glycan-protein interaction.

Therefore, an integrated systems approach to investigate structure-function relationships of glycan, or so called glycomics, is indispensable. For this reason, international collaborative efforts such as the Consortium for Functional Glycomics are resulting in the development of novel resources and technologies for glycomics. In the next subsection, the technologies and the bioinformatics platform discussed in the paper will be reviewed.

Concepts and Methods

Functional genetics approach to glycomics

By directly linking the role of glycosylation of proteins and glycan diversification to the phenotype at the cellular and the whole-organism level, it helps to understand how genotype influences the phenotype of the entire organism. For example, recent phenotype analysis of knockout strains of siayl and fucosyl transferases has revealed interesting phenotypes that provide evidence of specific glycan sequences in mediating aspects of cell-surface biology.

Nevertheless, to resolve how glycans modulate whole-organism phenotype, the functional genetics and whole-organism phenotyping studies should be coupled with measuring gene expression of glycan biosynthesis enzymes and their binding proteins which correlates with the glycan structures.

Development of glycol-gene microarray for glycomics

Using genome-wide arrays to investigate gene expression of enzymes involved in glycan biosynthesis and that of glycan binding proteins has certain limitation. For instances, the current human and mouse genome microarrays have limited representation of glycan biosynthesis enzymes and the sensitivity in measuring expression of these genes relative to other downstream events is limited. As a result, glyco-gene-based DNA microarrays focusing on glycan biosynthesis and binding protein genes were designed. These microarrays provide information on simultaneous expression of glycan biosynthetic enzymes that can be then correlated with the actual glycan structures.

Glycan analysis – from high-throughput to fine structure characterization

Characterization of the primary chemical structure of glycans is crucial to the study of the functions of glycans. To this end, several biochemical and analytical methodologies have been developed. For high-throughput analysis, there are methods like mass spectrometric (MS) and high-performance liquid chromatography (HPLC). For fine structure characterization, there are techniques such as nuclear magnetic resonance (NMR) and MS-MS. To enable a sensitive fine structure characterization with a high through-put analysis, informatics based sequencing methodologies that incorporate data from multiple complimentary techniques have been developed.

Biochemical analysis of specificity of glycan-protein interactions

To assess the relative binding affinities of glycan binding proteins (GBPs) and for designing inhibitors to physiological glycan-GBP interactions, chemical synthesis strategies like solid-phase synthesis have been employed to synthesize glycan structures that capture the glycans' diversity. Besides, lectin-based approaches have been used to fingerprint glycosylation on glycoproteins. And glycan arrays have been used to screen for novel ligand specificities for GBPs and for development of antibodies to target specific glycan motifs. Having the knowledge on ligand specificity of different GBPs, it helps to understand how cellular phenotype is modulated by glycol-related gene expression.

Bioinformatics platform for glycomics

In order to fully understand the structure-function relationships of glycans, it is inevitable to cut across multiple datasets. For this the paper proposed to use a bioinformatics platform to store, integrate and process the information generated by the aforementioned methods and disseminate it in a meaningful way through the Internet to the scientific community worldwide. In recent years, organizations like CFG are making substantial efforts to build databases such as Glyosuite database, KEGG Glycan database and tools for representation and analysis of glycan structures.

Integration of information or say data would be significantly facilitated by defining relationships between different entities. The specification of the conceptualization is called Ontology. The ontology that captures data definitions and inter-relationships in glycomics databases is quite complex due to their nature. Thus, the paper proposed to use an object-based relational database to capture complex relationships between the diverse and intrinsically hierarchical data. For hiding this complexity from the users during data acquisition and dissemination, it further proposed to develop an application with three-tier architecture which facilitates scientists to easily deposit data from and into the database, links orthogonal data sets derived from identical or similar samples for them to query the multiple datasets, and provides a portal to information and data ranging from molecule to mouse (see *Fig 1*). Last but not least, the bioinformatics platform was suggested to support computational tools to perform data mining analysis on the large scale glycomics data sets. Data exchange formats such as XML was also suggested for consistent description of glycan structures and glycomics data sets so as to set standards for incorporating glycan structures into a database to develop the glycan database into an international resource similar to GenBank and SwissPort.

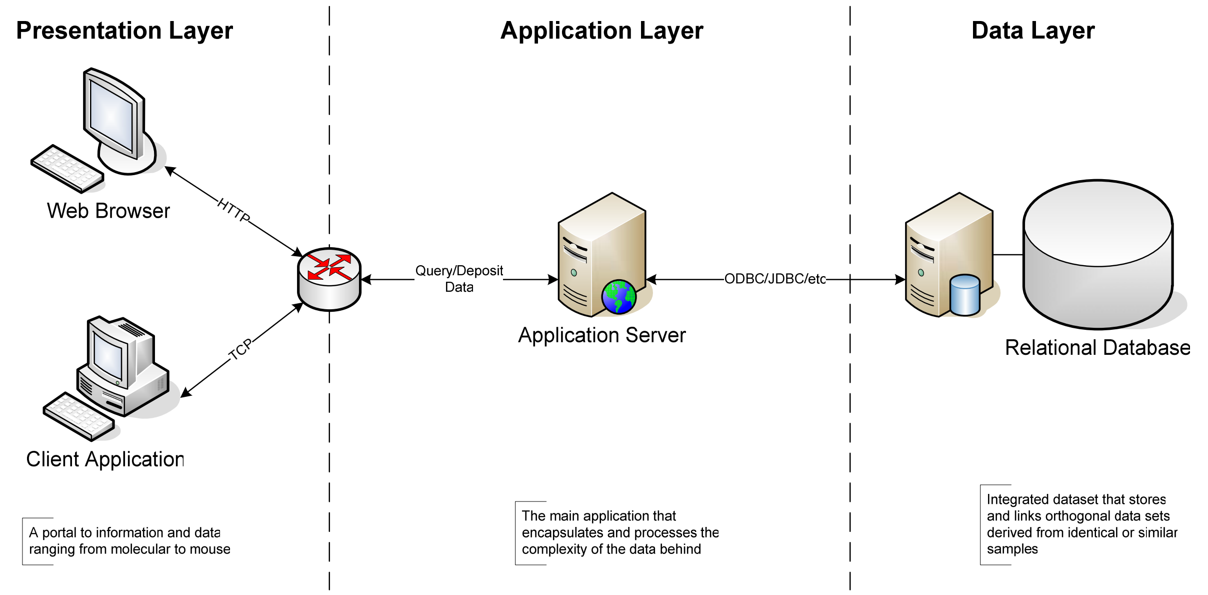
Presentation Layer
Application Layer
Data Layer
Web Browser
HTTP
TCP
Client Application
Query/Deposit Data
Application Server
ODBC/JDBC/etc
Relational Database
A portal to information and data ranging from molecular to mouse
The main application that encapsulates and processes the complexity of the data behind
Integrated dataset that stores and links orthogonal data sets derived from identical or similar samples

Figure 1 The 3-tier architecture of the bioinformatics platform for glycomics. The presentation layer is the user interface that communicates with the main application. The application layer is the middleware that communicates with the users, processes their requests with domain logic, and accesses the database accordingly. The data layer is where the back-end relational database situates. It stores the data and annotates their relationship.

Comments

It is clear that functional analysis, gene expression analysis, structural analysis, and biochemical interaction analysis of glycans are all essential to the understanding of glycans and its impacts on organisms. If we just look into one of these areas, we can hardly know how it affects an organism as a whole, predict their effects, and design new drugs for its related diseases. This problem not only exists in the study of glycans, but also in other biology fields like neuroscience. For example, there are large and rapidly growing quantities of heterogeneous neuronal data from many different levels (gene, synapse, neuron, pathway, etc) of research and of many different types (physiological, behavioral, image, etc)[10].

Given the fact that these kinds of data are all related in the same domain, the advantage of having an integrated platform which collaborates, correlates, or even analyses the diverse datasets is indisputable. Moreover, the hierarchical and sometimes reticular structure of the data relationships by no means can be fully and efficiently represented by traditional relational databases. Therefore, using an object relational database would be definitely an added advantage in representing the ontology of the domain concepts in glycomics or even other biology fields. However, object relational database is not the only solution to this issue. XML, RDF, and OWL are alternatives. More investigations should be carried out to carefully study which data structure would be most suitable for representing the data from glycomics.

A centralized international resource system like GenBank is a good approach but what is as important is to make use of the existing huge amount of data which have already been generated. This issue leads to the need of data format standardization. A standardized data format not only defines a consistent description of glycan structures and glycomics datasets across different large scale glycomics initiatives, but also facilitates the integration of data. After the data format is standardized, there will also be a need to develop some conversion tools to help biologists converting their existing data into the new format in order to get the data either automatically or manually published to the system. Finally, data interoperability would be the distributed issue that we ought to deal with. Although data exchange can be performed among different systems with a standardized data format, it is still necessary to develop or standardize a common protocol such that different systems, like data mining tools, can communicate with the centralized system, retrieve the relevant data and analyze them. Example of such kind of protocol could be Web Services.

Techniques in experimental biology and computational biology all add up could significantly advance the study of glycan structure-function relationships. A bioinformatics platform like the one mentioned here that bridges multiple datasets collected using different technologies would further facilitate the analysis of the data or enable us to find hidden facts among the data. As a consequence, the target audiences of this paper should not be limited to biologists. It should also be reviewed by people who are doing bioinformatics researches.

Section Two: Proposal

Introduction

Nowadays there are more than twenty public web-based resources for glycomics, not including those proprietary databases in glycan-related laboratories[11]. Examples of these kinds of resources are the CFG Glycan Database from USA, KEGG Glycan Database from Japan, Glycan NMR Profiles from Germany, and Lectin Database from UK. With the increasing number of glycan-related databases, there is an urgent need of a system to efficiently integrate these different databases and release the power of the data. One of the cores of CFG, namely Information and Bioinformatics Core (Core B), has been actively developing several complex relational databases and interfaces to facilitate linking appropriate data (from within the consortium as well as from other related databases worldwide) and engineering them for best usability.

This proposal will introduce several popular integration approaches, describe their advantages and disadvantages, and discuss a methodology and the methods of integrating glycan-related databases.

Integration Approaches

Application specific solutions

The most straight-forward and traditional way of integrating different databases with different schemas is to develop a special-purpose application which dedicates to solve a particular integration problem. The advantage of using this method is that the integration problem can always be solved since the application is usually hard-coded with a specific logic which is only applicable to that problem. Moreover, it oftentimes does not require any extra knowledge out of the current domain, such as the knowledge of external data. However, there are quite a lot of disadvantages. The application is usually not extensible. Whenever you need to integrate with more databases or systems, you need to change the application or even rewrite it[4]. Moreover, the application goes too complicated if integrating with large scale of databases.

Workflow systems

Workflow systems usually deal with exchange of electronic documents between different systems since manual flow of documents requires extra manual procedures and cannot ensure the documents are accurately flowed. It also makes sure the documents being exchanged are compatible to and reaching their corresponding and appropriate systems. Although they are more extensible than application specific solutions and provide support for routine results from one source to other sources, they still provide only limited help for comparing and manipulating data[4].

Data warehousing and data federation

A data warehouse integrates and stores relevant data from different independent sources into a central database or a database cluster in advance of the queries which depend on the diverse data. On the other hand, a data federation approach tries to build a composite view of the disparate and usually relational data that enables the users to query these data through the view. It will translate the query into local queries for retrieving the relevant data to fulfill the request[12]. All in all, the warehouse integration approach emphasizes data translation whereas the federated approach emphasizes query translation[10]. Using data warehouse the data will be translated according to a common schema and it makes the composition of query much easier and the data more manageable. Taking the data federation approach would keep the data up-to-date but it needs a great effort in translating the query into local queries. Even though they both provide a practical solution to data integration, they do not seem to be suitable in integrating data with complicated ontological and semantic relationships.

Semantic Web

"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation"[2,5]. It is a collaborative effort of W3C and a large number of researchers and industrial partners.

The Semantic Web provides a common framework that allows data, defined and linked in a way that it is machine-understandable, to be shared and reused across applications on the web[9]. There are currently two standards, the Resource Description Framework (RDF) and the Web Ontology Language (OWL), which are recommended by W3C.

RDF is an XML which integrates a variety of applications and provides a lightweight ontology system to support the exchange of knowledge on the web[3]. OWL is basically an extension of RDF. It expresses the meaning and semantics better than XML, RDF, and RDF schema since it has added more vocabulary for describing the classes (terms) and their relationships in the ontology[8]. Currently, there are three sublanguages of OWL. They are, in increasing expressive power, OWL-Lite, OWL-DL, and OWL-Full. Since OWL-Full is designed to support users who want to express their ontology to an extent that it is nearly impossible to have any computational reasoning on it, the reasoning software, or what we call reasoner, nowadays can only support OWL-Lite and OWL-DL.

A Semantic Web Platform

Methodology

As described in the previous section, a three-tier bioinformatics platform would probably be the most suitable way to solve the integration problem for glycomics or any other fields alike. However, instead of using an object relational database, applying a federation approach with web services[6] on the semantic web technology in the data layer would be even much more efficient (see *Fig 2*).

To integrate databases with heterogeneous schemas such as those glycan-related databases, semantic web can provide a framework that allows the existing data to be shared and reused. Those existing databases are not required to change their schemas in order to employ the new technology. The necessary procedure is to first define and standardize a set of ontologies, by using some ontology-engineering workbench to facilitate the construction of ontologies[7], for different datasets like those from glycan gene expression analysis and structural analysis. After that we need to map the existing data to the predefined ontologies and then use a conversion tool to convert the relational data according to the mapping into a semantic web language format on demand. Other tools, such as data mining tools, can also retrieve this standardized data for their own use. To facilitate the retrieval of data, web services can serve as the interface for both internal and external systems by receiving queries, passing it to the reasoner with the data, and returning the results back.

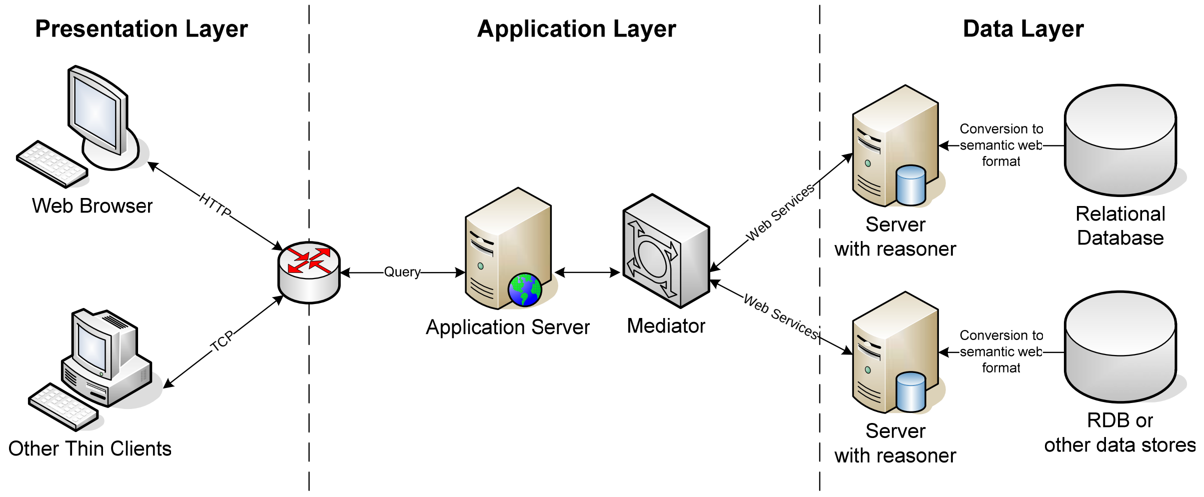

Figure 2 A semantic platform for glycomics. The presentation layer is the user interface that interacts with the users. The application layer is the middleware that receives queries from the users and then forwards it to the mediator behind. The mediator analyzes the query, translates it into sub-queries and then submits it through web services to different databases accordingly. The data layer is the distributed environment which consists of the disparate databases and is able to convert the data from the relational databases, or other data stores, into the semantic web format. The data are converted on demand based on the sub-queries and returned to the mediator after processed.

When the data from the individual databases are standardized and ready to be retrieved, a mediator shall be developed to unite the data. The mediator should be able to process a query, translate the query into sub-queries, gather the data returned from different databases, and then analyze it as a whole. Therefore, the mediator not only needs to support the semantic web language format, but also needs to be smart enough to break the query into sub-queries and reconstruct the segmented data. This kind of approach is somewhat similar to a data federation approach, but it is applied on semantic web and implemented by using a mediator.

While the mediator serves as a central unit responsible for processing the query from the user, a client interface shall exist for the sake of communication with the users. It gets the query from the users, posts it to the mediator, and displays back the result from the mediator to the users. A thin client interface, such as a web interface, would be most suitable since users from worldwide can connect to the system using their existing browsers.

Methods

The Web Ontology Language Description Logic (OWL-DL) will be used since it is the most expressive sublanguage of OWL which can be processed by a reasoner. Protégé, an ontology editor developed at Stanford, will be used to create the ontologies since it has a graphical user interface which aids the construction of OWL. D2RQ, a conversion tool developed in Germany which can convert relational data into RDF-like format according to an N3 mapping file, will be used to convert the data in the relational databases into OWL. And Jena, a Java API developed by HP, will be used to run the D2RQ tool and convert the data on the fly. A web service will be used to receive incoming requests with an ontology query and pass the query to Jena which in turn invokes D2RQ to get the relevant data in OWL format and also invokes the reasoner behind to analyze the data. Pellet, an open-source OWL reasoner, will be used. Although now Jena only supports the RDF query language, RDQL, changes can be made to Jena such that it can support more query languages.

The central unit of the whole platform is the mediator and the input for the mediator can be an ontology query language. RDQL and OWL-QL are suitable candidates. The mediator, as described before, will be developed to interpret the query and translate it into different sub-queries. It will also be able to combine the results from different databases and reconstruct them. Finally, a web interface which interacts with users will be developed in Java, a platform independent programming language. The web interface and the web services will be hosted by a Java web container such as Apache Tomcat.

References

1. Alberts, B. *et al.* "Molecular biology of the Cell." *Garland Science* G:16, 2002
2. Berners-Lee, T., Hendler, J. & Lassila, O. "The Semantic Web." *Scientific American* May 2001
3. Brickley, D. *et al.* "Resource Description Framework (RDF)." 04 Oct. 2005. *The World Wide Web Consortium (W3C).* 09 Dec. 2005 <http://www.w3.org/RDF/#specs>.
4. Haas, L. M., Lin, E. T. & Roth, M. A. "Data integration through database federation." *IBM Systems Journal* Vol 41, No. 4, 2002: 578-596.
5. Hendler, J., Berners-Lee, T. & Miller, E. "Integrating Applications on the Semantic Web." *Journal of the Institute of Electrical Engineers of Japan* Vol 122(10), Oct. 2002: 676-680.
6. Jhingran, A. D., Mattos, N. & Pirahesh H. "Information integration: A research agenda." *IBM Systems Journal* Vol 41, No. 4, 2002: 555-562.
7. Maedche, A. & Staab, S. "Ontology Learning for the Semantic Web." *IEEE Intelligent Systems* Mar./Apr. 2001: 72-79.
8. McGuinness, L. D. & Harmelen, V. F. "OWL Web Ontology Language Overview." 10 Feb. 2004. *The World Wide Web Consortium (W3C).* 09 Dec. 2005 <http://www.w3.org/TR/owl-features/>.
9. Miller, E. *et al.* "Semantic Web." 08 Dec. 2005. *The World Wide Web Consortium (W3C).* 09 Dec. 2005 < http://www.w3.org/2001/sw/>.
10. Miller, L. P., Wang, T., Liu, N.,Zhang Q., Cheung, K., Shepherd, G., Silberschatz, A., McDermott, D. & Marenco, L. "A Pilot Exploration of Ontology Mapping Issues among Neuronal Databases and Their Implications for Database Federation in the Neurosciences." *To be submitted.*
11. Raman, R., Raguram, C., Venkataraman, G., Paulson, J. C. & Sasisekharan, R. "Glycomics: an integrated systems approach to structure-functuion relationships of glycans." *Nature Methods* Vol. 2, No. 11, Nov. 2005: 817-824.
12. Stonebraker, M. "Too Much Middleware." *Sigmod Record* Vol. 31, No. 1, Mar. 2002: 97-106