Insights of New Tools In Glycomics Research

Denong Wang[1]* and Srinubabu Gedela[2,3]

[1]Stanford Tumor Glycome Laboratory, Stanford University School of Medicine,
Beckman Center, Rm B006, 279 Campus Drive, Stanford, CA 94305-5120, USA

[2]Center for Biotechnology & International Center for Bioinformatics, Andhra
University College of Engineering, Visakhapatnam-530003, India.

[3]Institute of Glycoproteomics & Systems Biology, Andhra Pradesh, India

*Corresponding authors: Denong Wang, Stanford Tumor Glycome Laboratory, Stanford University School of Medicine,
Beckman Center, Rm B006, 279 Campus Drive, Stanford, CA 94305-5120, USA, E-mail: dwang1@stanford.edu
Srinubabu Gedela: srinubabuau6@gmail.com

Since the origin of Journal of Proteomics & Bioinformatics the equivalence of papers published from different –omics disciplines is steadfast. The present editorial describes the new tools in glycomics research.

-omics era

Completion of the genome sequencing projects not only provides insight into the complex origin, history and relatedness of the species, but also helps in understanding molecular pathology of genetic diseases. Unreveling of the repertoire of genes in an organism opens the door to the universe of Proteomics and Glycomics diversities and to the new levels of understanding of the complexity of living species.

The "omics" are usually named based on its main molecular target, such as Genomics for the Genomes, Proteomics for the Proteomes, and Glycomics targets the universe of sugar chains and carbohydrate-containing macromolecules. There is, however, no boundary among "omics" in exploring the mysteries of lives. Instead, there is increasing need of integrated studies of living organisms. Systems Biology & Medicine emerged to accommodate with these new developments.

Glycomics is Complementary to Other –omics

Carbohydrates are prominently displayed on the surfaces of cells and present in many secretory proteins in bodily fluids. Expression of complex carbohydrates by human cells are characteristically associated with the stages or steps of embryonic development, cell differentiation, as well as transformation of normal cells to abnormally differentiated tumor cells. Sugar moieties are also abundantly expressed on the outer surfaces of the majority of viral, bacterial, protozoan and fungal pathogens. Many sugar structures are pathogen-specific, which makes them important molecular targets for pathogen recognition, diagnosis of infectious diseases, and vaccine development.

Altered glycosylations are associated with many human diseases. Structural and functional changes of glycans may thus serve as biomarkers of tumors, some type of cancers, peripheral vascular diseases, coronary artery diseases, inflammation, diabetes mellitus and associated complications (microvascular-nephropathy, retinopathy and neuropathy; macrovascular-stroke, cardiomyopathy and stroke). Augmented acquaintances on disease associated alterations in glycosylation tessellations and its data integration with proteome, genome and transcriptome endows new intrinsic biomedical insights and thus, far-reaching peradventures diagnostic application (preventions, early stage identification and cure) as well as for the new therapeutic discovery.

Identification of glycan markers for pathological conditions is, thus, one of the major jaunts for glycomics research. To let this area of postgenomics research with focus on new tools and technologies, scientists and scholars must be attracted, and they should be trained in this highly bustling area of research that crosses the barriers between various disciplines.

Emerging High-throughput Glycomics Tools

Biophysical, biochemical and classical immunological methods have proven very valuable in studying carbohydrate-carbohydrate and carbohydrate-protein interactions. They were, however, designed to monitor carbohydrate-based molecular recognition on a one-by-one basis and have limited analytical power or throughput in practical applications. In the past few years, a number of experimental approaches have been developed to fill this gap. Glycans profiling and sequencing with MALDI-MS, MALDI-FTMS, CID MS/MS and MALDI-TOF-MS are developed for high throughput structural characterization of glycomes. Advances in the development of microarray-based approaches are poised to accelerate progress towards better understanding of the roles of carbohydrates in biology and medicine. Carbohydrate microarrays, lectin & antibody arrays, neoglycolipid (NGL) technology coupled with mass spectrometry for carbohydrate ligand discovery, fluorescent neoglycoconjugate probes, and aminoglycoside antibiotic microarrays are among the many new tools becoming available to glycobiologists.

In spite of their technological differences, the microarray-based technologies are all solid phase binding assays for carbohydrates and their interaction with other biological molecules. They share a number of common characteristics and technical advantages. First, they contain the capacity to display a large panel of carbohydrates in a limited chip space. Second, each carbohydrate is spotted in an amount that is much smaller than that required for a conventional molecular or immunological assay. Thus, the bioarray platform makes effective use of carbohydrate substances. Third, they have high detection sensitivity. The microarray-based assays have higher detection sensitivity than most conventional molecular and immunological assays.   This characteristic is attributed to the fact that the binding of a molecule in a solution phase to an immobilized micro-spot of ligand in the solid phase has minimal reduction of the molar concentration of the molecule in solution. Therefore, in a microarray assay, it is much easier to have binding equilibrium take place and result is high sensitivity. Different platforms of microarrays are technically complementary and helpful for addressing many of problems in the glycomics research.

Consortium for Functional Glycomics

The Consortium for Functional Glycomics (CFG) has taken solid steps in facilitating key areas of glycomics research. One of the goals of CFG is to define paradigms by which glycan binding proteins mediate cell communication. Data towards achieving this goal is compiled in the CFG database, which can be accessed at <www.functionalglycomics.org>. Among the resources as services offered by the CFG to its investigators are glycan analysis, glyco-gene microarray, mouse knockout strains, mouse phenotype analysis, carbohydrate compounds, and glycan microarrays for screening the specificity of glycan binding proteins, including recent development of CFG pathogen carbohydrate arrays. A challenging issue is to select and obtain glycans of highest priority to the community since over 8800 structures are reported in the Bacterial Carbohydrate Structure Database (www.glyco.ac.ru/bcsdb/start.shtml).

Analogous to CFG in US, there is emerging of consortiums of glycobiologists in various countries worldwide: (http://www.functionalglycomics.org/static/consortium/links2Website.shtml)

Tumor Glycome Laboratories of the NIH Alliance of Glycobiologists for Detection of Cancer and Cancer Risk

Numerous studies comparing normal and tumor cells have shown that changes in glycan structures on the cell correlate with cancer development. Compared to molecular proteins, molecular glycans are extremely abundant and recent advances in technology have now allowed the effective systematic study of these structures. In late 2007, the National Cancer Institute launched a new initiative to discover, develop, and clinically validate cancer biomarkers based on complex carbohydrate structures attached to proteins and lipids. (http://prevention.cancer.gov/programs-resources/groups/cb/programs/glycome).

Seven Tumor Glycome Laboratories have been funded to search for glycan-based biomarkers for breast, ovarian, lung, prostate, and pancreatic cancers and melanoma. NCI's Tumor Glycome Laboratories are the principal component of the new trans-NIH Alliance of Glycobiologists for Detection of Cancer and cancer Risk. The other members of the Alliance are the Consortium for Functional Glycomics supported by the National Institute of General Medical Sciences and the Glycomics and Glycotechnology Resource Centers supported by the National Center for Research Resources.

Integrating Data

The challenge for -omics is to tackle the problem of fragmentation of knowledge by integrating the many sources of

heterogeneous information into a systematic entity. It is widely recognized that successful data integration is one of the keys to improve productivity for stored data. Through proper data annotation tools and algorithms, researchers may correlate relationships that enable them to make better and faster decisions about new inventions. The need for data integration is essential for present glycomics community, because glycomics data is currently spread across geographically in wide variety of formats. These formats can be integrated and migrated across platforms through different techniques. One of the important and widely usable, easily understandable techniques is the XML.

Extensible Markup Language

Today, there is an augmented requirement in different – omics fields to combine available software tools into chains, thus building multifarious applications from existing single-task tools. To create such workflows, the tools involved have to be able to work with each other's data - therefore, a common set of well-defined data formats is needed. Opportunities for new ways of combining and re-using data are arising as a result of the increasing use of web protocols to transmit structured data where Extensible Markup Language (XML) is coming into the picture; The XML is a general-purpose markup language helps in sharing, migrating data across heterogeneous systems and storing information, this facilitates data integration and application through the adoption of standards for representing certain types of data, e.g., genome, proteome, glycome, metabolome annotations. XML is designed to provide a document markup language that is easier to learn, pain less retrieve, comfortable to store, effort less transmit, and semantically richer than HTML. It provides a common format for expressing both data structures and contents making it a standard for data representation and transformation. Hence, it can help in integrating structured, semi structured, and unstructured data over the web repositories.

Computational Biology and Informatics Laboratory at the University of Pennsylvania came forward for managing and analyzing genome sequences using Extensible Markup Language Genomics Unified Schema (XMLGUS). XMLGUS facilitates amalgamating data into GUS by (i) formulating an XML interface that includes relational database key constraint definitions, (ii) regularizing traversal through that XML, (iii) realizing automatic processing of the XML with database key constraints and (iv) allowing for special processing of input data within the framework for spontaneous processing. Web data such as nucleotide and protein sequences are retrieved by agent programs, stored in the extensible markup language (XML) files, then parsed and loaded into

the secondary database, and transformed into the hyper text markup language (HTML) files for publishing. Once a standard is agreed upon, all databases and applications that store or process the data can share a common interface. It has major advantages are its ease of use, wide support available from software, database and LIMS (Laboratory Information Management System) vendors, and the large number of tools that exist to facilitate its use. A steadily growing number of databases, software applications and tools are XML-compliant, such as NCBI BLAST, DOAJ, PIR, SWISS-PROT, InterPro etc.

Database Federation

A federated database is a logical association of independent databases that provides a single, integrated, coherent view of all resources in the federation. It is one approach to data integration in which middleware, consisting of a relational database management system, provides uniform access to a number of heterogeneous data sources. Developments in our ability to integrate and analyze data held in existing heterogeneous data resources can lead to an increase in our understanding of biological function at all levels. However, supporting ad hoc queries across multiple data resources and correlating data retrieved from these is much more complex. To handle this conditions mediator like P/FDM helps, which integrates access to heterogeneous distributed biological databases. Our architecture makes use of the existing search capabilities and indexes of the underlying databases, without infringing on their autonomy. This technique leverages the native data management and search capabilities of individual source databases and creates a single, unified, logical view of the federated databases. Two major problems in constructing data federations (for example, data warehouses and database federations) concern achieving and maintaining consistency and a uniform representation of the data on the global level of the federation. The first step in creating uniform representations of data is known as data extraction, whereas data reconciliation is concerned with resolving data inconsistencies. In a proxy cache for federations of scientific databases it is important to estimate the size of a query before making a caching decision. With accurate estimates, near-optimal cache performance can be obtained.

Even though the most common means of data integration can be done with programming model, such as CORBA(Common Object Request Broker Architecture), J2EE (Java 2 Platform, Enterprise Edition), that access sources of interest directly and combine the data retrieved from those sources with the application itself. This approach always works, but it is expensive. Database federation is

cost effective and it does not require modification of the primary data stores as most are large number of heterogeneous databases to deal with. Furthermore, many databases (i.e., PDB) are in the public domain and thus not directly modifiable by researchers.

Controlled Vocabularies

Controlled vocabularies are integrating heterogeneous glycomics data sources are based on one of a common field, ontology or cross-reference. In a glycomics context, controlled vocabularies offer a form of data integration by enforcing naming conventions for data elements that ultimately appear in glycomics databases.

Statements and Recommendations

New Tools and Technologies

A number of high-throughput glycomics tools have reached or are very close to the stage of the current nucleic acid-based microarrays that are readily available for practical uses. Technical issues that require immediate attention may include but are not limited to optimization of existing technologies for array construction, quality control and technical standardization in both microarray production and application, establishment of specialized bioinformatic tools to handle the massive amount of carbohydrate microarray data and to effectively extract diagnostic or research information from each microarray assay.

Exploring the repertoires of glyco-epitopes and their receptors represents a long-term goal of carbohydrate research. How big is the repertoire of glyco-epitopes? Addressing this question is one of the most important topics in the post-genomics era. It was estimated that there are about 500 endogenous glyco-epitopes in mammals. However, this estimation did not consider the repertoires of the "hybrid" structures that are generated by protein posttranslational modification, including both *N*- and *O*-glycosylation. Further more, the conformational diversity of carbohydrates and micro-heterogeneity of carbohydrate chains substantially increases the repertoire of carbohydrate-based antigenic determinants or glyco-epitopes. Considering carbohydrate structures of the microbial world, which are directly relevant to medicine, the sizes and diversity of the repertoires of glyco-epitopes are unpredictable. Establishment of high-throughput platforms of carbohydrate microarrays provides powerful means to facilitate the identification and characterization of carbohydrate-based pathogen signatures and other biomarkers.

Joint effort by academic and industrial sectors is highly recommended to direct the establishment of libraries of monoclonal antibodies, lectins and other carbohydrate-binding proteins. These biomolecules are critical for defining glyco-epitopes and are useful for detection of glyco-epitopes in living organisms. Thus, using specific immunological probes to characterize glyco-epitopes is equally important to the structural determination of glyco-epitopes. Similar effort has been successfully made for protein-based biomarkers. A notable example is the establishment of a large collection of monoclonal antibodies for cell differentiation antigens (CD antigens). Availability of specific probes for CD antigens, in combination with the state-of-art technologies of flow cytometry (Hi-D FACS), has revolutionized research in cellular biology and immunology and medical applications of CD antigens, especially in the clinical diagnosis of leukemia and other human diseases. Exploring the repertoires of glyco-epitopes and their cellular receptors, with the aid of new Glycomics tools and specific immunological probes, represents one of the highly active areas of postgenomics research that may last for a few decades and likely accompanied with a fruitful outcome.

Mass Spectrometry and Its Prone

Mass spectrometry (MS) is appearing as an empowering technology in different fields. Coupling capillary electrophoresis, hydrophilic interaction chromatography and other advanced miniaturized separation techniques with this emerging technology-is the hub of present proteomics, metablomics and glycomics research.

Protein glycosylation play crucial roles in preservation of the function of a glycoprotein. Therefore, prominent streamlined sample preparation protocols are prerequisite in preparation of a biological active protein product, including protein-based drugs. Glycans profiling and sequencing with MALDI-MS, MALDI-FTMS, CID MS/MS and MALDI-TOF-MS are the most impressive approaches for high throughput glycomics applications. During the experimentation, the production of vast amount of glycomics MS data is a large-dimensional optimization problem. Threshold gradient descent regularization algorithms, numerical simulation studies, systematic representation of glycan structure may be suitable for theoretical generation of all possible first and second generation fragments resulting from glycosidic and cross ring cleavages. Despite the wealth of data provided by the most overwhelming MS and MS/MS techniques, most of them are not reproducible. There is a need to develop suitable validation parameters to get robust results.

Mathematical Modeling-Inter Disciplinary Approach

The increasing use of mathematical modeling in –omics research is inevitable as biology becomes more quantitative. The complexity of the biological sciences assembles inter disciplinary involvement requisite. For the software engineer, biology opens up new and exciting branches while for the biologist, mathematical modeling offers another research appliance launches with a new powerful laboratory technique but only if used appropriately and its limitations identified.

Few models, N-Glycosylation, galactosylation of an oligosaccharide are mathematically formulated. There is a requirement to develop new model identifies the substrate specificities of known glycoprotein. These models are additionally extended to encompass enzyme kinetics, biomarker prediction, structure annotation and sequence analysis using nonlinear algebra. However, the use of esoteric mathematics arrogantly applied to biological problems by software engineers who know little about real biology, together with unsubstantiated claims as to how important such theories are, also little to promote the interdisciplinary involvement which is so essential.

Databases and Required Improvements

The challenge for present glycomics community is to develop existing and upcoming information accessible from beginning to end in one master database system. Today, however, GlycomeDB database available with the linking of the glycan data in the seven major databases such as Bacterial carbohydrate structure database (BCSDB), Complex carbohydrate structure database (CCSD), consortium for functional glycomics (CFG), Glyco Base(Dublin), Glyco Base(Lille), GLYCOSCIENCES.de, Kyoto encyclopedia of genes and genomes (KEGG). However, it will be useful to provide integrated data (preferably XML format) instead of linking. This database should contain 1) a single glycan structure that collide with well circumscribe reliance standard, 2) The earliest known data for the structure, combining literature reference, an account of its biological origin, and its analytical information used in its project. The master system has to collide with the intrinsic complexity of glycomics research and open up the information (preferably open access) for the scientific community. Additionally, the new database should be cross connected with genomics, proteomics, metablomics, protein structure databases, and it should have robust annotation tools.