

Subproject 4: Bioinformatics of Glycan Expression

William S. York, Senior Investigator, and Amit Sheth, John Miller and Krzysztof J. Kochut, Investigators

Summary

Glycomics is an emerging discipline, which still lags behind proteomics in terms of the development of bioinformatic tools that are required to track and process vast amounts of raw data, make it accessible to scientists with diverse backgrounds, and deduce important but non-obvious data relationships that can be interpreted in the context of developmental and pathological states of human cells. Our approach to this challenge is the development of an integrated data system that includes workflow protocols and tools for keeping track of experimental samples and processes, data processing tools to extract relevant information from the raw data, database schema to save the resulting data, and ontological tools that will facilitate access to the information and reveal systematic relationships within the data collected here, as well as among diverse data that is distributed in databases throughout the world and within the domain knowledge of the ontology itself. The basic design of this system has been developed, placing highest priority on the interoperability of its component parts. To this end we are currently developing two ontologies, GlycO and ProPreO. GlycO incorporates knowledge of glycan structure, function, biosynthesis, and metabolism. ProPreO incorporates knowledge of proteomic analysis and the resulting experimental data. These ontologies thus describe fundamental relationships between glycomics concepts and their association to experimental data, allowing individual elements of the data to be classified and viewed in the overall context of the biological/biochemical system. These ontologies will serve as the glue that ties the components of our bioinformatics system together and as a semantic basis for a portal that we will develop to facilitate data access and to reveal relationships within the data. A key component of this portal will be a graphical browsing and querying interface that we are developing. The highly integrated nature of our bioinformatics system for glycomics is a prerequisite for its optimal functionality, with each component being designed such that its format and content are consistent with the GlycO and ProPreO ontologies.

Ontology Development and Data Exchange

To make knowledge machine accessible, it must be formalized. One formal representation of knowledge is an ontology, *i.e.*, "a specification of a conceptualization that is designed for reuse across multiple applications." An ontology thus embodies a formal encoding of the concepts, relations, objects, and constraints within a semantic model of a domain. (P.D.Karp, et al., <http://www.ai.sri.com/~pkarp/xol/xol.html>). The three flavors of the Web Ontology Language (OWL) provides a means to build ontologies with a good compromise between expressiveness and computational complexity on one hand and versatility and simplicity on the other. We are using OWL to develop a suite of ontologies for the glycomics domain: the Glycan Ontology (GlycO) embodies knowledge of the structure and metabolism of glycans; the Proteomics Process Ontology (ProPreO) embodies knowledge of experimental and computational processes that are used to generate and interpret glycoproteomic data; a third ontology, which is not yet implemented, will embody knowledge of the biological functions and interactions of glycans.

Currently, the GlycO ontology contains over 550 classes that correspond to concepts that describe the structural features of glycans. One of these concepts is the *carbohydrate_residue*, or basic unit of glycan structure. Carbohydrate residues are classified according to their structural features, such as absolute conformation (D or L), overall configuration (e.g., *gluco* or *manno*), anomeric configuration (α or β), ring form (*f* or *p*), and number of main-chain carbons (e.g., hexosyl or pentosyl). Thus, the concepts in GlycO can be mapped to language commonly used by the glycobiochemist to describe the building blocks of glycans. By formalizing the specification of glycosyl linkages between carbohydrate residues, GlycO also provides a means to represent the chemical environment of specific instances of these residues. GlycO implements a powerful extension of this approach by defining "canonical" residue instances, as described in the following example. A typical *N*-glycan contains a single β -D-Man_p residue in its core. This residue is glycosidically linked to a specific site (oxygen-4) of the next residue, which is invariably a β -D-Glc_pNAc residue. The identity of the β -D-Man_p residue and its precise location in the core of the *N*-glycan allow it to be unambiguously classified. In

fact, glyco biologists often refer to this residue as “the core β -Man residue”, with the implied assertion that this residue is in a particular molecular location and that its biosynthetic addition to the glycan was catalyzed by a specific class of glycosyl transferases (*i.e.*, a GDP-mannose-dolichol diphosphochitobiose mannosyltransferase, EC 2.4.1.142). The trained glyco biologist can intuitively make a large number of structural and biochemical inferences when the core β -Man residue is invoked. This can be viewed as a colloquial classification of a canonical glycosyl residue, as each unique *N*-glycan structure contains a single glycosyl residue called “the core β -Man residue.” However, very few of the residues that make up *N*-glycans have a common name based on their chemical identity and context.

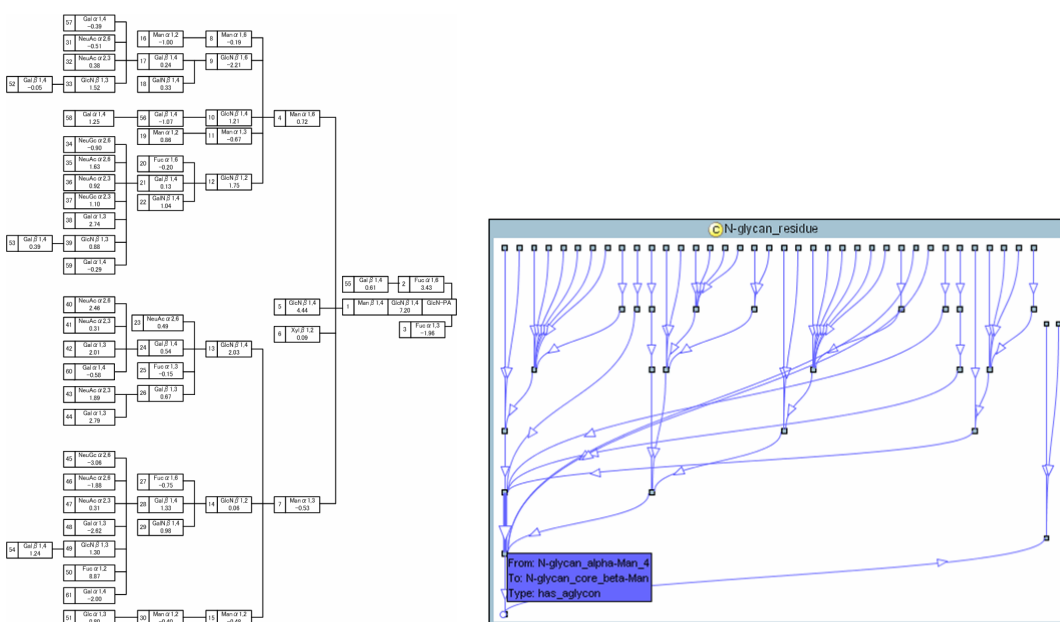


Figure 1. Formalization of the canonical GlycoTree (left) in the Glyco ontology (right).

Specification of canonical residues in Glyco extends this powerful concept to all of the monosaccharide residues within the glycan. For *N*-glycans, this is accomplished by defining a canonical tree that subsumes all *N*-glycans. That is, almost all known *N*-glycans can be completely specified by choosing a subset of the nodes of this canonical tree that form a connected (directed) graph. Such a graph (known as glycoTree) has been previously described (N. Takahashi and K. Kato, *Trends Glycosci. Glycotech.*, **15**: 235-251), and we have formalized that structure as a collection of interconnected, canonical residue instances in Glyco (Figure 1). This provides a mechanism by which the chemical

and biological properties of each residue within the glycan, as well as the cellular machinery involved in its biosynthesis and degradation, can be semantically inferred. That is, other semantically defined objects (such as glycosyl transferases) and processes (such as metastasis) can be associated with canonical residues that they interact with or depend on. Some of these associations may be indirect (via other objects in the ontology), or inferred by analysis of quantitative information (e.g. correlation of the abundance of glycans containing a specific canonical residue and the observation of a cellular property like invasiveness) that could be extracted from a semantically annotated database. An example is specification (within Glyco) that addition of “N-glycan_b-D-GlcNAc_9” is catalyzed by an instance of the GNT-V class of glycosyl transferases, and that glycans containing this residue is present are recognized by the lectin LPHA. Then, the hypothesis that GNT-V overexpression is correlated with elevated invasiveness of various types of cancer cells can be inferred from a semantically annotated database that includes information regarding the binding of different lectins to various cancer cell lines and the physiological properties of these cell lines.

Automated protocols are being used to populate Glyco. In order to harvest this data, we use the Semagix Freedom toolkit that allows extraction of data from semi-structured internet sources, such as Carbohydrate Bank, KEGG and SweetDB. Simply collecting this information is not enough, since database schemas are usually shallow and categorization is typically done by keywords rather than by a class hierarchy. Keywords rarely provide sufficient information for the complete classification of glycan instances after extraction from the source. For incorporation into Glyco, instances of glycans *and their constituent residues* have to be classified according to their structure. This process is facilitated by first converting the imported glycan structure (usually in IUPAC format) into the LINUCS format (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), and then to our GLYDE format. Both LINUCS and GLYDE are tree-based formats in which the natural topology of the glycan is mirrored in the data structure. GLYDE differs from LINUCS in that GLYDE is an XML-based format that can be readily parsed by widely available software. GLYDE-encoded glycan instance information can be parsed according to a canonical tree, such as GlycoTree embodied within Glyco (see

above). In this process, the glycan is split up into its component residues and each residue is categorized according to its chemical structure and context.

In addition to its compatibility with Glyco, GLYDE is a highly expressive and generally useful format for data exchange. We have therefore developed utilities to convert various common formats for glycan structure into GLYDE format and are exposing these as web pages (HTML) and Web services. We invoke the IUPAC to LINUCS web service provided by SweetDB as part of this process.

ProPreO, part of the ontology suite, is being developed as a comprehensive process ontology that models experimental proteomics. The current ontology metrics for ProPreO feature 290 classes and 189 slots; including facets and instances add up to a total of 495 frames. ProPreO categorizes the phases of proteomics experiment into three major sections (reference: PEDRO UML Schema, <http://pedro.man.ac.uk/uml.html>): (a) Separation technique, (b) Mass Spectrometry and (c) Data Analysis. It also includes concepts relating to data types (including parameter lists for chromatography and mass spectrometer runs, raw and processed analytical data) and quantification of chemical constituents used in the experiments. The ontology incorporates five fundamental concepts we evolutionarily derived to form a fundamental framework to describe any proteomics experiment; namely (i) data, (ii) data_processing analysis, (iii) hardware, (iv) parameter list and (v) task. These classes are defined in an hierarchical manner (*mass_spectrometer* is a subclass of *instrument*), in which sibling classes are distinguished by logical restrictions and the definition of class-specific properties. For example, a *mass_spectrometer* is defined as *hardware* -> *instrument* -> *mass_spectrometer*. A large collection of mass spectrometer instruments are included in ProPreO. A particular class of mass spectrometer like '*ABI_Voyager_DE_Pro_MALDI_TOF_mass_spectrometer*' has the property '*has_source*', which is qualified as '*Has a MALDI source as its source*'. Thus, each mass spectrometer model is defined as a class in the ontology and each actual mass spectrometer is specified as an instance whose properties correspond to its components and configuration.

ProPreO forms the basis of our research initiative in *semantic annotation of scientific data*, that is, generation of metadata for diverse classes of experimental proteomics data. Semantically annotated data specifies critical syntactic, provenance and contextual information. The annotation process has two aspects. The first is addition of semantic metadata to the actual data file, making it possible for software that is aware of the ProPreO ontology to understand the content and context of the data. The second is recording instances of the metadata in the ontology itself, which constitutes a knowledgebase that is “aware” of the general content, context, and location of various instances of actual data. Ultimately, we see the ProPreO ontology (in conjunction with the GlycO ontology) as the centerpiece of a system that will allow the automated annotation, retrieval, and semantically valid comparison of widely distributed data. For example, this system would allow mass spectral data that embodies information regarding the abundance of particular N-glycan structures at the surface of different cell types to be found even if the data were collected in different laboratories. Data retrieval would involve examination of metadata instances that describe (1) properties of the analyte sample, such as its biological source (e.g., cell type) and chemical class (e.g., permethylated glycan) and (2) the type of data (e.g. LC-MS data). The suitability of the designated data for quantitative comparisons would be judged by analysis of metadata that specify, for example, the type of instrument used and the ionization method employed. A key part of this system would be the inclusion of metadata that describes sample history, so that differences in sample preparation and handling could be taken into account.

Stargate - Web Service based Portal and Visualization Tools

As a novel approach towards information integration for distributed glycomics/glycobiology information and data we have developed a Web Service based web portal – Stargate. We aim to make key features of Stargate available as downloadable tools, to be used by research groups in the glycobiology domain. The application features the following subsystems:

a) BioUDDI (*BUDDI*) – Using the Universal Description, Discovery and Integration (UDDI) protocol of Web Services, BUDDI is a dedicated registry for listing all available Web Services in glycobiology domain. Unlike generic UDDI maintained by Microsoft or IBM, BUDDI classifies Web Services using life sciences taxonomy. This is an intuitive approach for *publishing* and *searching* Web Services under relevant categories. We aim to develop this as a worldwide community resource.

b) Format Converter – We have linked the various available glycan representation format to the xml-based glycan representation standard (GLYDE). This utility, using a combination of services available at SweetDB and the CCRC, converts given IUPAC representation to LINUCS and LINUCS to GLYDE format.

c) Group Forum – A forum providing the members of the research group to post messages, post papers, schedule meeting and upload/download research files.

Other standard features of a web portal, such as authentication and search facilities (for published papers, posters etc.), will also be included.

Ultimately, the core technology of this portal will be a graphical interface for representation, browsing, and querying of the GlycO and ProPreO ontologies. Such a visualization tool is required because ontology representation languages are not designed for human readability. Our ontologies, which are built using the Web Ontology Language (OWL), are internally represented as graphs composed of nodes (representing classes and instances) and edges (representing relationships between the nodes). This form naturally lends itself to visual representations. However, we expect that, when mature, GlycO and ProPreO will embody thousands of classes at the schema level and perhaps millions of instances. Such large ontologies pose major technical challenges, as tools used for their visualization should be capable of transforming them into representations that can be intuitively interpreted by biologists and other users. Based on our examination of these technical challenges (described below) and the technology that can address them, and have developed an overall design for the ontology visualization tool, which we call OntoVista.

The visualization tool should be capable of loading both instance and schema files and have options to efficiently switch between them. It must overcome the display limitation (the number of nodes that can be displayed at one time) while providing efficient methods to navigate the entire ontology schema/instance graph. It should provide the user with an efficient means to browse through the ontology using a query interface. It should provide options to generate simplified representations of the (potentially huge) instance graph by filtering unwanted instances and/or relationships. Finally, it should provide links to external resources that provide additional information regarding specific classes, instances, and relationships.

We have examined the currently available tools and have found that none of the tools contain all of the following features, which are included in our current design of GlycoVista :

1. Display OWL schema along with restrictions (including those inherited from ancestor classes).
2. Provide options to filter out unwanted properties, restrictions, and inherited properties.
3. Graphically display the results of a semantic search based on concept or instance descriptions in the ontology, providing the user with an efficient means to browse through the ontology using a query interface.
4. Using a technique such as MapView to aid navigation through large graphs.
5. Use node clustering techniques to simplify the instance graph. For example, each glycan, which is composed of its constituent residues, could be rendered as a single node that represents a cluster of residue nodes, simplifying the visualization of relationships among different glycans.
6. Display only instances that have common relationships. In combination with features 2 and 5, this could be used to reduce the number of instances in the graph and to generate specific, refined views for the biologist. For example, displaying only those nodes that correspond to glycan instances that are related by the property *is_precursor_of* and the enzymes (*e.g.*, glycosyl transferases) that are related to these glycans by the property *has_substrate* constitutes a method to render a metabolic pathway.

Data Processing and Workflow

Glycoproteomics analysis involves the collection, storage, processing, and retrieval of large amounts of data. The identification and quantitation of glycans, glycopeptides and glycoproteins is most often performed using mass spectrometry, due to its high sensitivity and capacity to provide structural information. However, a considerable amount of sample preparation and molecular separation is also required for effective glycoproteomics analysis. Ultimately, the success of this process depends on development in two areas: (1) the ability to perform individual tasks in a high-throughput, repetitive fashion, and (2) the ability to automatically coordinate and keep track of individual tasks and to initiate those tasks, such as large-scale data processing and annotation, that can be performed without human intervention.

Fortunately, development of many of the protocols and algorithms included in the first area has been actively pursued by analytical equipment manufacturers and other researchers. However, it is still necessary to develop modules that facilitate data exchange between the commercially available tools and provide additional functionality, such as data filtering, visualization, and parameter extraction. To address these requirements, we have developed several utilities for the processing and visualization of LC-MS and LC-MS/MS data, with emphasis on the analysis of glycopeptide identification. For example, the MassLynx software (Micromass) has utilities to process LC-MS/MS data to generate peak list files that can be directly used by the Mascot software for MS/MS ion-searching. This is a powerful method of identifying peptides based on their tandem MS fragmentation patterns. Unfortunately, the peak lists generated by the MassLynx utility do not contain any LC retention time (RT) information, which is a critical parameter that can be used for other purposes, such as quantitation of peptides and mapping their chromatographic properties. We developed a utility to map individual records in the peak list file to data in the original raw data, thereby extracting the RTs corresponding to each query processed by Mascot. The Mascot output can then be annotated to indicate the RTs of peptides that are identified. This makes it possible to identify peptides in LC-MS data recorded under similar chromatographic conditions, using the mass and RT data in the annotated Mascot output files. Raw LC-MS data can

be visualized using a set of programs we have developed using the C and R programming languages. These programs use binning to generate a data matrix from which one- and two-dimensional graphical representations can be easily extracted. This matrix can also be used for peptide mapping and quantitation, provided that information regarding the identity, mass, and retention times of peptides listed in Mascot output files are available, as described above.

We have also developed a utility for preprocessing protein sequence databases to simplify ion-searching and filtering (e.g., via Mascot) for glycopeptide identification. The input for this utility is a collection of FASTA protein sequence records, in which amino acid residues are represented by single letters. The utility replaces all occurrences of the letter N (asparagine residues) in the N-glycosylation consensus sequence (N-X-S/T) with J, a letter not used for any amino acid. Effective MS/MS ion searching of data for a collection of glycopeptides that have been deglycosylated with PNGase F can then be accomplished by executing Mascot with the modified database as input and the instruction that J represents an amino acid that has the same mass as aspartic acid (D), the amino acid residue that is generated upon deglycosylation with this enzyme. This approach is superior to the standard approach (instructing Mascot to include the possibility of deamination of asparagine residues in its analysis) because it is faster and lowers the false discovery rate by specifically distinguishing deglycosylated glycopeptides from peptides that were deaminated by other mechanisms or peptides that were misidentified as deaminated due to inappropriate selection of isotopomer peaks as precursor ions. These (actually or artifactually) deaminated peptides are not included in the Mascot output when the protocol that uses the modified database is implemented. In combination with the size-exclusion chromatography (SEC) method for glycopeptide enrichment (developed within the Core 2 project), this protocol provides two criteria for glycopeptide identification. The first criterion (the presence of an N-glycosylation consensus sequence, corresponding to J in the peptide sequence) is trivial to recognize in the output file. The second criterion (a peptide mass less than the SEC cutoff used for glycopeptide enrichment) provides confirmatory evidence to identify deglycosylated glycopeptides. The second criterion cannot be strictly applied, as some deglycosylated glycopeptides will still have masses above the SEC cutoff.

The second area of development required for high-throughput glycoproteomic analysis includes automated workflow protocols to coordinate the entire process and initiate those tasks that can be done without human intervention. Workflow protocols define how tasks are structured, who performs them, what their relative order is, how they are synchronized, how information flows to support the tasks, and how tasks are being tracked. We formally specify laboratory workflow in XML Process Definition Language (XPDL) using Enhydra JaWE (Java Workflow Editor), an open source graphical Java workflow process editor designed to Workflow Management Coalition (WfMC) specifications. The workflow (formalized in XPDL) can then be executed using Enhydra Shark, an extendable workflow engine framework also based on WfMC specifications. Shark can be configured to exploit the organizational structure defined on a lightweight directory access protocol (LDAP) server, using almost any DB system for storing information. For execution of system activities, Shark uses the WfMC "ToolAgents" API, which facilitates the development of agents that can execute native system applications, java classes, java script, Web service operations, email operations, etc.

We are using these tools to implement preliminary workflow protocols for the glycopeptide identification process described above. These will be extended to more complicated workflows that include parallel chromatography steps and comparative quantitation schemes as they mature in the laboratory. Ultimately, the workflow protocols will include automatic annotation of data using concepts defined in the GlycO and ProPreo ontologies, facilitating subsequent access and interpretation via semantic tools.