Presented at the W3C Workshop on Semantic Web for Life Sciences, Cambridge MA October 27-28, 2004.

Semantic Web technology in support of Bioinformatics for Glycan Expression[1]

Amit Sheth, William York, Christopher Thomas, Meenakshi Nagarajan, John A. Miller, Krys Kochut, Satya S. Sahoo, Xiaochuan Yi
Large Scale Distributed Information Systems (LSDIS) lab and Complex Carbohydrate Research Center (CCRC), the University of Georgia
http://lsdis.cs.uga.edu/Projects/Glycomics/

Due to the complexity of biological systems, interpretation of data obtained by a single experimental approach can often be interpreted only if viewed from a broader context, taking into account the information obtained by many diverse techniques. The vast amount of interpreted experimental data that is now available *via* the internet opens the possibility of collecting the relevant pieces of information that will enable scientists to form hypotheses based on the integration of this diverse information. However, the sheer volume of data that is available makes it very difficult to select the information necessary to make a coherent model of the biological system under study. We are developing an integrated semantic methodology to address this challenge, with the current emphasis of supporting bioinformatics applications in glycomics. Our ontology-driven methodology starts with the development of a domain ontology (or interrelated domain ontologies) and associated models for supporting trust and provenance, and leads to semantic search, browsing, and integration, further aiding in analysis of experimental data and heterogeneous documents.

GlycO (for "Glycomics Ontology") is being populated with extensive domain knowledge that embodies semantically rich descriptions of carbohydrate structures, glycan binding relationships, glycan biosynthetic pathways, and the developmental biology of stem cells. Classes of objects and their relationships in GlycO model information we store about the differential expression of glycan structures on the surface of developing stem cells. We are developing methods to automate the population of these ontologies from multiple, heterogeneous (semi-structured and structured) knowledge sources. For example, the structure ontology is populated with specific glycan structures and the building blocks (glycosyl residues) from which these molecules are assembled. Provenance, i.e., the sources of the information and the processing history of curated and non-curated data, is also incorporated into the ontology description and knowledge base. Data provenance records about derived data reflect not only when and how but also why certain derivations have been made. Provenance information on how to reproduce experimental results increases re-usability and reproducibility of data. Information of this kind helps to establish trust in the data and the ensuing analysis.

For the formal representation of our model we use the Web Ontology Language (OWL). Even though other formalisms for knowledge representation exist and are more expressive, OWL presents a good trade-off between expressiveness and time-complexity of the inference/reasoning procedures. Furthermore, it is becoming the standard for ontology representation on the web. For populating the ontology, we use the Semagix Freedom, a commercial product resulting from earlier research and technology transfer from the LSDIS lab [1]. We aggregate knowledge from multiple sources such as SweetDB, KEGG, GO, UMLS and others. Since the contents of the sources overlap in many cases, we disambiguate the entities before adding them to the ontology. One advantage of drawing the instances from multiple sources is greater "fault tolerance" or reinforcement. Another is the ability to aggregate information about one instance from multiple sources. While one source might contain a more complete description of the chemical properties of a glycan, another might contain more information about its relationships to other biochemical substances. A toolkit being developed in this project also includes tools for visualizing and browsing the ontology, forming and executing

[1] This work is funded primarily by *Bioinformatics of Glycan Expression (one of the four components of the "Integrated Technology Resource for Biomedical Glycomics," funded by the* National Institute of Health, July 1, 2003 – June 30, 2008; http://sweet.ccrc.uga.edu/glycomics/glycomics.php).

meaningful semantic queries, and annotating databases by reference to ontological classes. The toolkit also allows automatically updating provenance information while populating the ontology and in future will support propagating provenance information while querying over the ontology.

The first step towards interactive querying of the ontology is to build a browser applet that gives the user several ways to access information in the ontology (Fig.1). The default view of the ontology is a tree view similar to that of Protege[2]. Also available is a TouchGraph[3] representation of the ontology. The different frames and tabs are logically interconnected. When the user clicks on a class in the tree view, it comes into focus in the graph view and vice versa. At the same time, the properties tab and the instances tab are updated with the properties and instances of the selected class. The natural language description and provenance information of a particular class or instance is always displayed in a text area at the bottom of the applet. Overall, the tool combines the advantages of both a hierarchical structure, provided by the tree view, and a graph structure, provided by the TouchGraph™ interface.

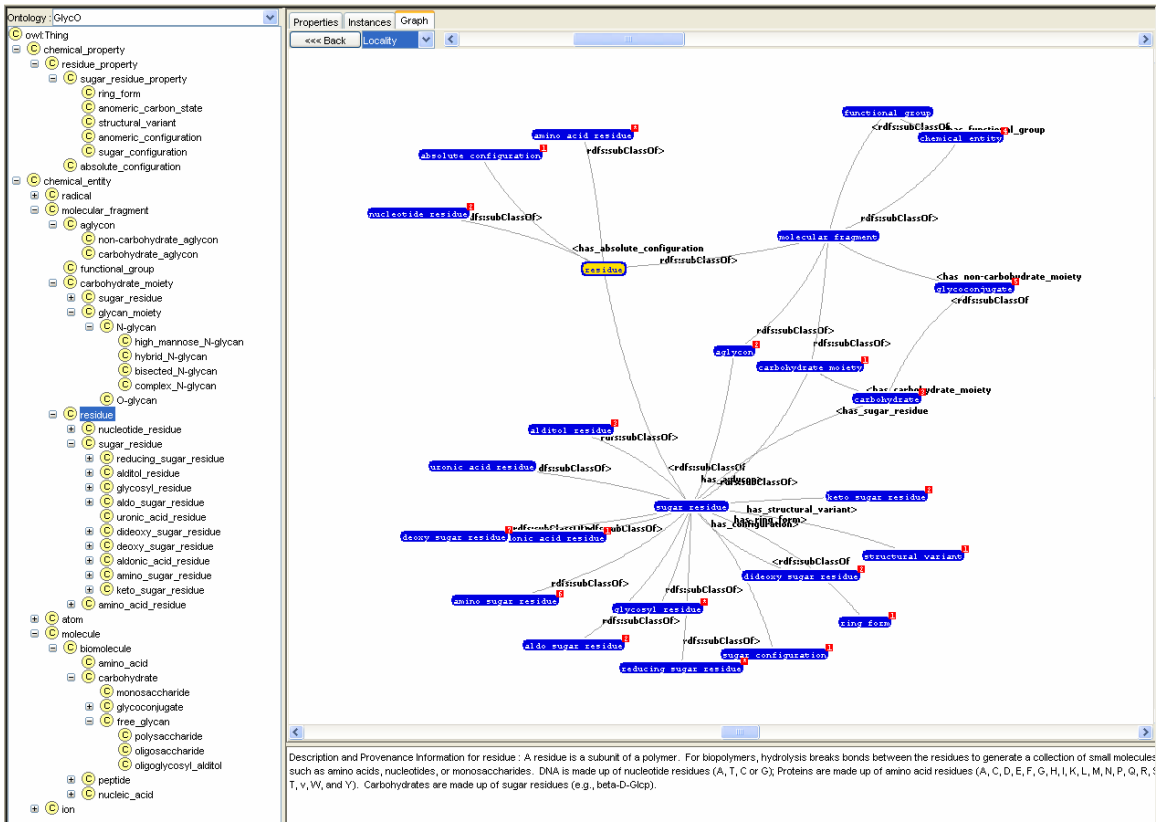

Figure 1: Ontology Visualizer showing tree view and graph view of GlycO

[2] http://protege.stanford.edu/
[3] http://www.touchgraph.com

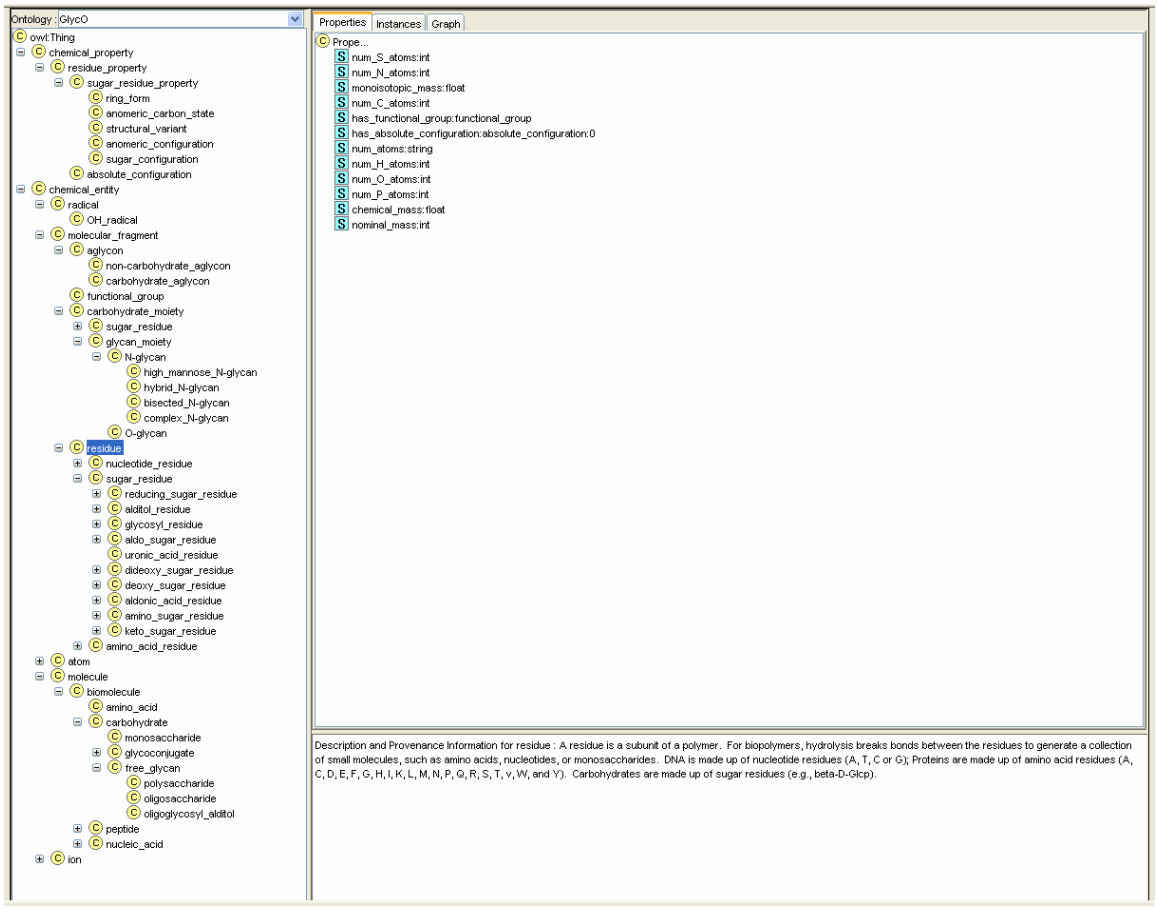

Figure 2: Ontology Visualizer showing properties of the selected class

Upcoming version of the ontology visualizer will support semantic search and browsing capabilities. Using the graphical representation of the ontology, we will be able to form complex semantic queries by selecting nodes and edges from the TouchGraph representation, by selecting classes and properties from the tree view representation, by typing a query in a form with contextually relevant fields, and such semantic querying and browsing will provide access to relevant knowledge in the populated ontology, annotation of experimental data or heterogeneous documents (i.e., metadata) or the related raw experimental data and relevant scientific publications and report. Concurrently, the domain discussed above is being extended with the development of an inter-related ontology to support proteome analysis.

[1] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, Y. Warke, Managing Semantic Content for the Web, IEEE Internet Computing, July/August 2002, pp. 80-87.

[2] A. Sheth and C. Ramakrishnan, "Semantic (Web) Technology In Action: Ontology Driven Information Systems for Search, Integration and Analysis," In *IEEE Data Engineering Bulletin, Special issue on Making the Semantic Web Real, December 2003, pp. 40-48*

For additional details, see project page: http://lsdis.cs.uga.edu/Projects/Glycomics/