Semantic Integration of Glycomics Data and Information

WS York[1], A Sheth[2], K Kochut[2], JA Miller[2],
C Thomas[2], K Gomadam[2], X Yi[2], S. Sahoo[2], and M Nagarajan[2]

[1]Complex Carbohydrate Research and Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA - USA
[2]Large-scale Distributed Information Systems Laboratory and Department of Computer Science, University of Georgia, Athens, GA - USA

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 diverse information that will enable scientists to form broadly based hypotheses. 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, utilizing an ontology driven process for glycomics as the application domain. This approach involves development of a set of interdependent ontologies, called GlycO (for "a Glycomics Ontology"). GlycO is being populated with extensive domain knowledge that embodies semantically rich descriptions of carbohydrate structure, glycan binding relationships, glycan biosynthetic pathways, and the developmental biology of stem cells. Classes of objects and their relationships in GlycO model information that 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, heterogenous (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 this information and the reliability of those sources, is also incorporated into the ontology description and the knowledge base. This system will include tools for visualizing and browsing the ontology, forming and executing meaningful semantic queries, and annotating databases by reference to ontological classes. Together, these semantic structures, metadata, and tools constitute the basis for a transition from a database model to a knowledgebase model for exploring and interpreting complex biological data, demonstrating a transformation from information retrieval for human analysis to interactive and automated knowledge discovery.