ProPreO - An ontology for high-throughput glycoproteomics

Satya S. Sahoo1,2, James A. Atwood III2, Cory Henson1, Amit Sheth1, Ron Orlando2, William S. York2

1. Large Scale Distributed Information System (LSDIS) Lab, Computer Science Department, 2. Complex Carbohydrate Research Center

Motivation
An ontology is a formal model of a domain. An ontology primarily consists of:

1. Structure of ProPreO
2. Semantic Provenance – tracks information to comprehensively describe the various stages of glycoproteomics experiment and relations between these techniques used in the generation of experiment datasets
3. Store, modify and retrieve experimental data - in an automated manner (without rate-limiting human intervention)

The three top-level concepts of ProPreO are:
1. Data - Data can be experimental (measured) or theoretical (calculated).
2. Material continuant – a real world object namely, instruments, biological or chemical agents.
3. Tasks - A process that is initiated or implemented by an agent

Introdution
An ontology is a formal model of a domain. An ontology primarily consists of:

a) Conceptual: These represent the generic classes of entities in the domain of interest. E.g. peptide
b) Relationships: The concepts are associated with each other by different types of relations. E.g. has_parent_protein human_fut8 [fucosyl transferase 8]
ProPreO is a process ontology that models the complete experiment lifecycle of glycoproteomics from cell culture to identification of (glyco)peptide.

ProPreO derives its power from the real-world instances that are used to create instance of the concepts defined in schema of the ontology. Real world entities are extracted from:
• Databases of experimental data
• Medical literature
• Web pages

The populated ProPreO ontology will be used to derive pertinent information using the experimental datasets.
ProPreO ontology population metrics:
1. Total Number of real-world Instances – 3 million
2. Total number of assertions (triples) – 19 million

Acknowledgement: ProPreO was developed as part of the Integrated Technology Resource for Biomedical Glycomics (5 P41 RR018502), funded by the National Institutes of Health, National Center for Research Resources.