

# ProPreO - An ontology for high-throughput glycoproteomics

Satya S. Sahoo<sup>1,2</sup>, James A. Atwood III<sup>2</sup>, Cory Henson<sup>1</sup>, Amit Sheth<sup>1</sup>, Ron Orlando<sup>2</sup>, William S. York<sup>2</sup>

1. Large Scale Distributed Information System (LSDIS) Lab, Computer Science Department, 2. Complex Carbohydrate Research Center  
The University of Georgia, Athens GA, <http://lsdis.cs.uga.edu/projects/glycomics/propreo/>



Complex Carbohydrate Research Center  
The University of Georgia

## 1. Structure of ProPreO

The structure of ProPreO ontology consists of concepts required to comprehensively describe the various stages of glycoproteomics experiment and relations between these concepts.

The three top-level concepts of ProPreO are:

- Data - Data can be experimental (measured) or theoretical (calculated).
- Material continuant – a real world object namely, instruments, biological or chemical agents
- Tasks - A process that is initiated or implemented by an agent

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

- Concepts:** These represent the generic classes of entities in the domain of interest. E.g. *peptide*
- Relationships:** The concepts are associated with each other by different types of relations. E.g. **TLILESQNRW** *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.

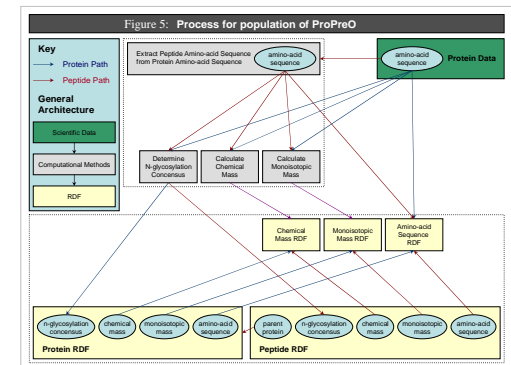
### Motivation

- Semantic Provenance** – tracks information regarding the procedures and instrumental techniques used in the generation of experiment datasets
- Semantic annotation of experimental glycoproteomics data** – enables the interpretation of experimental data by software applications
- Store, modify and retrieve experimental data** - in an automated manner (without rate-limiting human intervention)

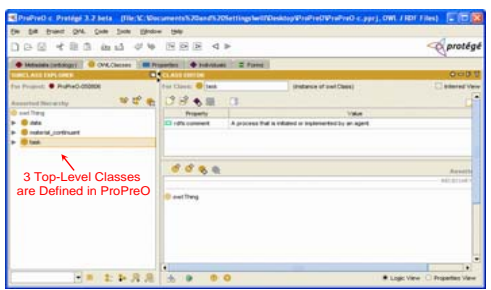
## 2. Population of ProPreO

An ontology 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

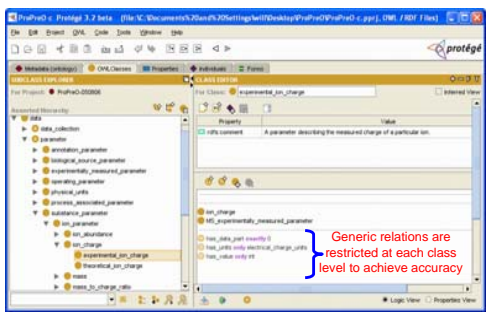


The ProPreO Ontology in Protégé



The ProPreO ontology has rich set of relations and attendant restrictions on them. This enables ProPreO to accurately model the concepts and relations between them.

Figure 2: The ProPreO Ontology in Protégé



## 3. Application: Semantic Annotation of Experimental Data

- Software applications, that form the foundation of high-throughput experiment protocols, cannot 'understand' the experimental
- The experimental datasets are annotated with concepts defined and described in an ontology
- Using these semantic software applications, without human intervention, can process and analyze data
- Semantic annotation of experimental data also forms the foundation of the semantic provenance. Semantic provenance enables the computation of the correct context for the processing or interpretation of experimental data

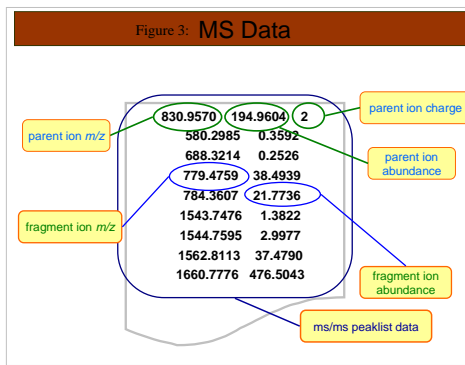


Figure 4: Semantic Annotation of MS Data

```
<ms/ms_peak_list>
<parameter
instrument="micromass_QTOF_2_quadropole_time_of_flight_mass_spectrometer"
mode="*msims">
  <parent_ion m/z = 830.9570 abundance = 94.9604 z=2/>
  <fragment_ion m/z = 580.2985 abundance = 0.3592/>
  <fragment_ion m/z = 688.3214 abundance = 0.2526/>
  <fragment_ion m/z = 779.4759 abundance = 38.4939/>
  <fragment_ion m/z = 784.3607 abundance = 21.7736/>
  <fragment_ion m/z = 1543.7476 abundance = 1.3822/>
  <fragment_ion m/z = 1544.7595 abundance = 2.9977/>
  <fragment_ion m/z = 1562.8113 abundance = 37.4790/>
  <fragment_ion m/z = 1660.7776 abundance = 476.5043/>
</ms/ms_peak_list>
```

The populated ProPreO ontology will be used to derive pertinent information using the experimental datasets.

- ProPreO ontology population metrics:
- Total Number of real-world Instances – 3 million
  - Total number of assertions (triples) – 19 million

