Knowledge Modeling and its Application in Life Sciences: A Tale of two ontologies

Satya S. Sahoo, Chris Thomas, Amit P. Sheth, William S. York, Samir Tartir

Paper Presented at
15th International World Wide Web Conference, Edinburgh, Scotland
May 25, 2006

Bioinformatics for Glycan Expression
Integrated Technology Resource for Biomedical Glycomics
NCRR/NIH
Outline

• Background
• Ontology Structure
• Ontology Population: Knowledge base
• Ontology Size Measures
• Applications in Semantic Bioinformatics
• Conclusions
Background: glycomics

- Study of structure, function and quantity of ‘complex carbohydrate’ synthesized by an organism
- Carbohydrates added to basic protein structure - Glycosylation

Folded protein structure (schematic)
Outline

• Background
• Ontology Structure
• Ontology Population: Knowledge base
• Ontology Size Measures
• Applications in Semantic Bioinformatics
• Conclusions
Requirements from ontologies

- Storing, sharing of data + reasoning over biological data → logical rigor
- Expressive as well as decidable language → OWL-DL
- Incorporation of real world knowledge → ontology population
- Ensure amenability to alignment with existing bio-medical ontologies
• **Challenge** – model hundreds of thousands of complex carbohydrate entities
• But, the differences between the entities are small (E.g. just one component)

**GlycO ontology**

• How to model all the concepts but preclude redundancy → ensure *maintainability, scalability*
ProPreO ontology

• Two aspects of glycoproteomics:
  o *What is it?* → identification
  o *How much of it is there?* → quantification

• Heterogeneity in data generation process, instrumental parameters, formats

• Need data and process provenance → ontology-mediated provenance

• Hence, ProPreO models both the glycoproteomics experimental process and attendant data
**Ontology-mediated provenance**

<table>
<thead>
<tr>
<th>Parent Ion m/z</th>
<th>Fragment Ion m/z</th>
<th>Fragment Ion Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>830.9570</td>
<td>194.9604</td>
<td>2</td>
</tr>
<tr>
<td>580.2985</td>
<td></td>
<td>0.3592</td>
</tr>
<tr>
<td>688.3214</td>
<td></td>
<td>0.2526</td>
</tr>
<tr>
<td>779.4759</td>
<td></td>
<td>38.4939</td>
</tr>
<tr>
<td>784.3607</td>
<td></td>
<td>21.7736</td>
</tr>
<tr>
<td>1543.7476</td>
<td></td>
<td>1.3822</td>
</tr>
<tr>
<td>1544.7595</td>
<td></td>
<td>2.9977</td>
</tr>
<tr>
<td>1562.8113</td>
<td></td>
<td>37.4790</td>
</tr>
<tr>
<td>1660.7776</td>
<td></td>
<td>476.5043</td>
</tr>
</tbody>
</table>

**Mass Spectrometry (MS) Data**

**Ontology-mediated provenance**
Ontology-mediated provenance

<ms-ms_peak_list>
  <parameter instrument="micromass_QTOF_2_quadropole_time_of_flight_mass_spectrometer" mode="ms-ms"/>
  <parent_ion m-z="830.9570" abundance="194.9604" 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"/>
  </parent_ion>
</ms-ms_peak_list>
Compatibility with existing Biomedical ontologies

- Top level classes are modeled according to the Basic Formal Ontology (BFO) approach
- Taxonomy of relationships and multiple restrictions per class → accuracy
- Hence, both GlycO and ProPreO are compatible with ontologies that follow BFO approach
- Exploring alignment with ontologies listed at Open Biomedical Ontologies (OBO)
Outline

- Background
- Ontology Structure
- Ontology Population: Knowledge base
- Ontology Size Measures
- Applications in Semantic Bioinformatics
- Conclusions
GlycO population

- Multiple data sources used in populating the ontology
  - KEGG - Kyoto Encyclopedia of Genes and Genomes
  - SWEETDB
  - CARBANK Database
- Each data source has different schema for storing data
- There is significant overlap of instances in the data sources
- Hence, entity disambiguation and a common representational format are needed
Has CarbBank ID?

IUPAC to LINUCS

LINUCS to GLYDE

Compare to Knowledge Base

Already in KB?

YES: next Instance

Semagix Freedom knowledge extractor

GlycO population

[][

Asn

{(4+1)[b-D-GlcpNAc]
{(4+1)[b-D-GlcpNAc]
{(4+1)[b-D-Manp]
{(3+1)[a-D-Manp]
{(2+1)[b-D-GlcpNAc]
{(4+1)[b-D-GlcpNAc]
{(6+1)[a-D-Manp]
{(2+1)[b-D-GlcpNAc]}}}

NO

Insert into KB

YES

LOGOS to GLYDE

YS

Compare to Knowledge Base
Has CarbBank ID?

IUPAC to LINUCS
LINUCS to GLYDE
Compare to Knowledge Base

YES
NO

Semagix Freedom knowledge extractor

Instance

Data

YES:

next Instance

Insert into KB

NO

<GlycO population>
ProPreO population: transformation to rdf
ProPreO population: transformation to rdf

Computational Methods
- Extract Peptide Amino-acid Sequence from Protein Amino-acid Sequence
- Determine N-glycosylation Consensus
- Calculate Chemical Mass
- Calculate Monoisotopic Mass

Scientific Data
- Protein Data
- Amino-acid Sequence RDF
- Monoisotopic Mass RDF
- Chemical Mass RDF

RDF
- "Protein RDF"
- "Peptide RDF"
- n-glycosylation consensus
- chemical mass
- monoisotopic mass
- amino-acid sequence

ProPreO population: transformation to rdf

Key
- Protein Path
- Peptide Path

ProPreO population: transformation to rdf

Large Scale Distributed Information Systems
Outline

• Background
• Ontology Structure
• Ontology Population: Knowledge base
• Ontology Size Measures
• Applications in Semantic Bioinformatics
• Conclusions
# Measures of ontology size

<table>
<thead>
<tr>
<th></th>
<th>GlycO</th>
<th>ProPreO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>318</td>
<td>390</td>
</tr>
<tr>
<td>Properties</td>
<td>82</td>
<td>32</td>
</tr>
<tr>
<td>(datatype &amp; object)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Property restrictions</td>
<td>333</td>
<td>172</td>
</tr>
<tr>
<td>instances</td>
<td>737</td>
<td>3.1 million</td>
</tr>
<tr>
<td>assertions</td>
<td>19,893</td>
<td>18.6 million</td>
</tr>
</tbody>
</table>
Outline

• Background
• Ontology Structure
• Ontology Population: Knowledge base
• Ontology Size Measures
• Applications in Semantic Bioinformatics
• Conclusions
Pathways do not need to be explicitly defined in GlycO. The residue-, glycan-, enzyme- and reaction descriptions contain all the knowledge necessary to infer pathways.
The N-Glycan with KEGG ID 00015 is the substrate to the reaction R05987, which is catalyzed by an enzyme of the class EC 2.4.1.145. The product of this reaction is the Glycan with KEGG ID 00020. Reaction R05987 catalyzed by enzyme 2.4.1.145 adds_glycosyl_residue N-glycan_b-D-GlcpNAc_13.
Semantic Web Process to incorporate provenance

Biological Sample Analysis by MS/MS → Raw Data to Standard Format → Data Pre-process → DB Search (Mascot/Sequest) → Results Post-process (ProValt)

Semantic Annotation Applications

Storage

Biological Information
Overview - integrated semantic information system

- Formalized domain knowledge is in ontologies
- Data is annotated using concepts from the ontologies
- Semantic annotations enable identification and extraction of relevant information
- Relationships allow discovery of knowledge that is implicit in the data
Outline

• Background
• Ontology Structure
• Ontology Population: Knowledge base
• Ontology Size Measures
• Applications in Semantic Bioinformatics
• Conclusions
Conclusions

• GlycO uses simple ‘canonical’ entities to build complex structures thereby avoids redundancy → ensures maintainability and scalability

• ProPreO is the first comprehensive ontology for data and process provenance in glycoproteomics

• Web process for entity disambiguation and common representational format → populated ontology from disparate data sources

• The two ontologies are among the largest populated ontologies in life sciences
Data, ontologies, more publications at Biomedical Glycomics project web site:
http://lsdis.cs.uga.edu/projects/glycomics/

Thank You