Leveraging an Alignment between two Large Ontologies: FMA and GO

John H. Gennari & Adam Silberfein
{gennari, ads99}@u.washington.edu
Biomedical & Health Informatics and The Information School,
University of Washington, Seattle WA

A long-term research objective of our group is to develop methods for connecting, aligning, or mapping disparate ontologies. In pursuit of this goal, we have found the technology of the Protégé system to be invaluable, both as a means of viewing and working with ontologies, and as an architecture on which to explore and experiment with new ideas and prototypes. In this abstract, we describe early work in pursuit of ontology alignment.

Our approach has been to work with a specific pair of ontologies that have a partial overlap, and try to envision the value and benefit of developing an alignment between these ontologies. Thus, our work so far is not focused on *how* to align these ontologies, but rather, to demonstrate the benefits that follow from a successful alignment.

Ontologies are sometimes used as a type of *schema* for underlying databases. One benefit from aligning disparate ontologies is that users can then view an external database through the lens of a locally-developed ontology. We have built a prototype that demonstrates this capability, via the use of a special slot-widget plug-in that allows for queries to remote, but aligned databases.

The Ontologies

The Foundational Model of Anatomy (FMA) is a large, comprehensive ontology for the symbolic modeling of the structure of the human body. It is called *foundational* as anatomy is the foundation for all biomedical domains. The FMA is designed to be a "reusable and general resource of deep anatomical knowledge which can be filtered to meet the needs of any knowledge-based application" (Rosse & Mejino, 2003). Thus, a primary functionality of the FMA is to provide a consistent, disciplined definition of the structural aspects of anatomy.

The Gene Ontology (GO) is a structured, controlled vocabulary to allow molecular biologist to better share data and knowledge about the roles of gene products (Gene Ontology Corsortium, 2001). The GO is developed *collaboratively* and has become a successful standard for annotation of genes and gene products. It is organized into three hierarchies of terms: (1) the molecular function of a gene product, (2) the larger-scale biological process that may involve the gene product, and (3) the cellular component that is important or relevant for the gene product.

An important aspect of the GO is that there exists a set of databases that contain information about particular gene products in particular research species that have been studied and annotated with the GO controlled vocabulary. For example, historically, the GO was developed by the groups developing *Flybase* (for drosophilia), the *Saccharomyces* Genome Database (SGD, for yeast), and the Mouse Genome Database (MGD). Current the GO web pages list more than 30 such "annotated" databases contributed by about 15 groups worldwide. The critical functionality that the GO provides is that researchers can use the GO terms to retrieve related gene products across these multiple databases.

The Alignment

The FMA and the GO overlap in a relatively minor way: they both include descriptions and terms for the components (or structure) of the cell. However, they differ considerably in the organization of those concepts. This should come as no surprise, given the vastly different approaches of the two groups, and given the raison d'etre for the two ontologies. Our goal was to demonstrate the value of connecting these two groups, by aligning their ontologies. Thus, on one side, we would hope that anatomists and those used to an anatomic organization of knowledge would be able to access the vast and growing databases of annotations made available through the GO consortium. For the other side, we would hope that molecular biologists might be interested in understanding how their proteomics and genomics research results might play out in a larger-scale anatomic view (perhaps even beyond the level of the cell).

The GO cell component hierarchy contains about 1400 terms, or less than one-tenth of the total number of GO terms. The FMA contains about 75,000 terms, but there are only about 1000 terms that describe the cell and its components. Our initial goal is to demonstrate the value of aligning ontologies, rather than on formulating new methods for ontology alignment. Thus, with the assistance of anatomists and cell biologists, we hand-mapped about 150 terms between the two ontology. The great majority of these mappings were straight one-to-one connections, but we did find a few instances of many-to-one maps in both directions: sets of FMA terms matching a single GO term, and sets of GO terms matching to a single FMA term.

Extracting a Cell-centric Partition of the FMA

Since our goal is to show the benefit of aligning the FMA and the GO, we assume our users are primarily interested in those concepts that are relevant for cellular biology. Thus, we have created a special version of the FMA that retains its overall structure and organization, but excludes all anatomical objects that are not related to the Cell and its components. More specifically, we have extracted an ontology that contains only the FMA classes Cell, Cell part, Cell Junction, Cell cavity, Cell substance, Macromolecule, and all of the subclasses of these.

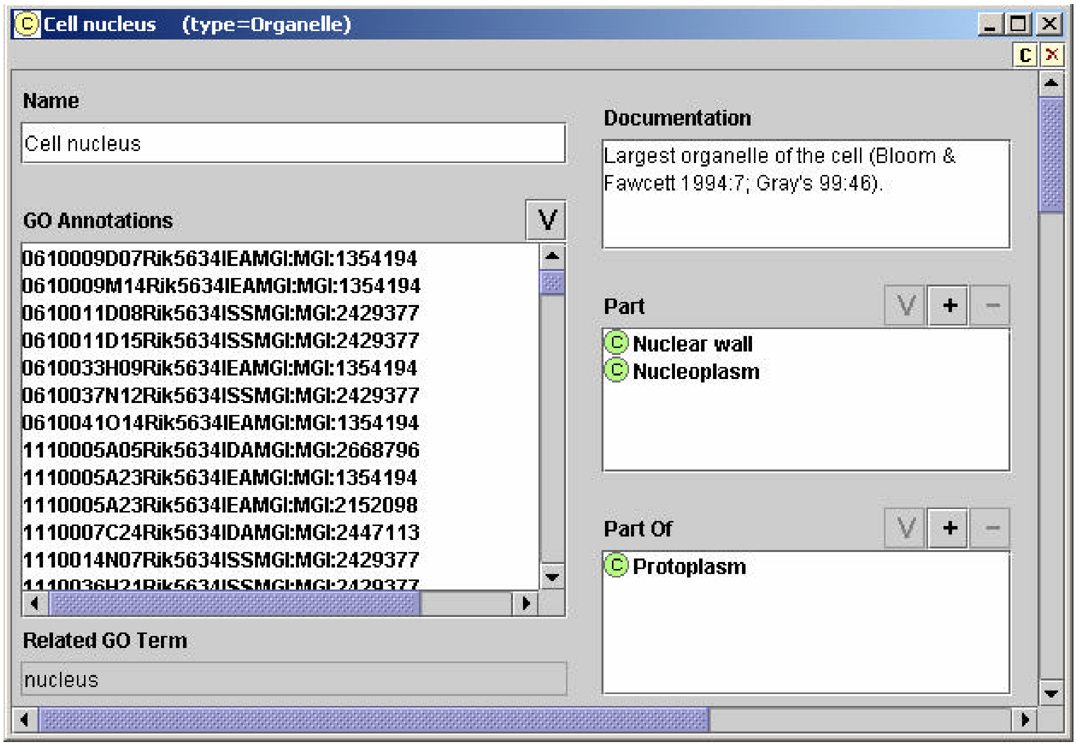
We have used the Prompt tool (a Protégé tab plug-in) to extract this version of the FMA, thereby creating a separate "cell biology" version of the FMA which contains only about 1000 concepts. (We are aware that this approach could lead to version control problems. If one extracts many such partitions of the FMA, how can the FMA developers control the evolution of their ontology? A better approach might be to build "non-materialized views" of the ontology, but this capability is not currently available.)

Connecting an Ontology to a Database

Given our alignment between concepts in the "cell biology FMA" (CB-FMA) and the cell component portion of the GO, our next step is to directly link the CB-FMA to the GO annotation databases. We carried this out by adding a new slot to classes in the CB-FMA called "GO annotations", and by associating that slot with a custom-developed slot-widget that carries out a set of database queries against the GO annotation databases. Figure 1 shows this slot widget within the FMA class of "cell nucleus". The term "nucleus" as shown in the lower potion of the slot widget is the result of our mapping between the FMA and GO. The list of annotations is provided by retrieving data from the GO annotation databases. One significant feature of this slot widget is that the distinction between the database information and the ontology is blurred. We believe this distinction is sometimes arbitrary, and is not important for our users.

Thus, the "GO Annotations" slots looks like any other slot. However, the information displayed in Protégé (via our widget) is **not** stored in the Protégé ontology. In fact, from Protégé's perspective, that slot is always empty (it has no value). However, whenever our slot widget is active for an FMA concept, it uses the concept ID to build and execute an SQL query that (a) retrieves the mapped GO term, and (b) uses that GO term to retrieve a set of annotations from the GO databases. The widget then displays these values as if they were Protégé slot values. If there is more than one GO term mapped to this particular FMA term, then the widget uses all of these terms, and displays the union of the sets of returned GO annotations.

As a further step, the GO annotation databases include a set of information about each annotation. Our slot widget allows users to view this information in a separate pop-up window via the "V" button (so as to be consistent with the Protégé look-and-feel). This includes information such as the database identifiers,



Cell nucleus (type=Organelle)

Name

Cell nucleus

GO Annotations

```
0610009D07Rik5634IEAMGI:MGI:1354194
0610009M14Rik5634IEAMGI:MGI:1354194
0610011D08Rik5634ISSMGI:MGI:2429377
0610011D15Rik5634ISSMGI:MGI:2429377
0610033H09Rik5634IEAMGI:MGI:1354194
0610037N12Rik5634ISSMGI:MGI:2429377
0610041O14Rik5634IEAMGI:MGI:1354194
1110005A05Rik5634IDAMGI:MGI:2668796
1110005A23Rik5634IEAMGI:MGI:1354194
1110005A23Rik5634IEAMGI:MGI:2152098
1110007C24Rik5634IDAMGI:MGI:2447113
1110014N07Rik5634ISSMGI:MGI:2429377
1110036H21Rik5634ISSMGI:MGI:2429377
```

Related GO Term

nucleus

Documentation

Largest organelle of the cell (Bloom & Fawcett 1994:7; Gray's 99:46).

Part

Nuclear wall
Nucleoplasm

Part Of

Protoplasm

Fig. 1: The database slot widget for the FMA term "Cell Nucleus". The related GO term ("nucleus") and the instances of GO annotations shown on the left are the result of SQL database queries.

the strength of evidence for this particular annotation, name of the gene product that is being annotated, and the date of annotation. Since none of this is information is stored within the Protégé ontology, it is not editable from the pop-up window.

Conclusions

We believe we have made a contribution in two directions. First, for many users the distinction between "data" and "knowledge" is fuzzy, unclear, and not important to their information-gathering tasks. By creating a plug-in that carries out SQL queries against a DB and embedding this plug-in within the default Protégé user interface, we allow users to seamlessly traverse back and forth between database information about gene annotations and knowledge base information about anatomy.

Second, by leveraging a set of (hand-crafted) mappings between the GO and the FMA, we believe we have enhanced the knowledge stored in both systems. Our next step is to evaluate this claim, by working with GO users and testing the value of our "anatomic" view of annotations.

Acknowledgements

We thank Jesse Wiley, Onard Mejino, and Augusto Agoncilla for domain expertise in cellular biology and anatomy. This work was funded in part by an NLM Bisti planning grant.

References

Rosse, C. and Mejino, J. L. V. (2003) A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics* **36**:478-500.

The Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. *Genome Research* **11**: 1425-1433. See also www.geneontology.org.