Combinatorial Approaches to Bio-Ontology Management With Large Partially Ordered Sets*

Cliff Joslyn† and Susan Mniszewski‡

Modeling, Algorithms, and Informatics (CCS-3)
Mail Stop B265, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Knowledge structures called "ontologies" (as an abuse of good philosophical terminology) have become increasingly important in modern knowledge systems. Characterized generally as "machine representations of a domain of knowledge", they have many different technical realizations, and ontological concepts have leaked into everything from simple function typologies to database schemata to object-oriented data structures to full-up agent-based AI. Famous general ontologies include Cyc and Wordnet; ontological capability forms one of the foundations of the nascent Semantic Web through technologies such as DAML+OIL [2] and OWL [9]; and ontologies are becoming increasingly important in knowledge-based sciences.

Despite recent excitement, ontologies have had relatively limitted penetration into mainstream computer science and systems. An exception to this is computational biology, where ontological concepts have escaped the laboratory and come into their own. The computational biological revolution has meant that for some biologists, the use of large databases, including such ontological structures such as the Gene Ontology (GO) [1], the Enzyme Commission Database (EC) [3], the MEdical Subject Headings (MeSH) [8], and others, is now a standard part of the work day.

To the extent that there is a common technological realization of ontological systems, they can be cast as large, taxonomically organized data objects over which automated inference and reasoning (for example using description logics) is performed. But what distinguishes the succesful ontologies in computational biology is that they are of an intermediate level of technical complexity. By avoiding the use of the inference engines, and instead existing as large collections of hierarchical, taxonomic, categorizations of biological objects such as genes and proteins, they come closer to being specially structured databases.

Most significantly for us here, such bio-ontologies are well cast as large relational objects, which then begs questions about their combinatorial properties. In particular, our fundamental thesis is that bio-ontologies such as the GO, the EC, and MeSH, share with object-oriented type hierarchies the formal structure of a **multi-poset:** a union of partially ordered sets (posets), each one representing a different semantic relation.

The GO in particular is a collection of three disjoint sets of **categories** (biological process, molecular function, and cellular location), with **annotations** assigning genes and proteins to sets of these categories. Each set of categories is then equipped with one partial order representing `is-a` inheritance, and another representing `has-part` composition, so that more general categories are towards the top, and more specific categories are towards the bottom. As shown in Table 1, the GO posets are relatively large objects, and more importantly, continue to grow exponentially as their collective authors continue to update them.

The computational biology community has recently seen work by groups using the GO for various tasks such as drug discovery, protein function inference, and automated annotation. One of the most fundamental such tasks is that of **categorization:** given a large list (on the order of hundreds) of genes of interest, analyze their distribution over the GO, and decide which node(s) best "summarize" the gene list. This then begs questions such as: are the genes concentrated in one place? in multiple places? if not, how dispered as they? what value should be placed on general vs. specific posible categorizations? and how is that measured?

We found existing work in this area insufficient, as it either did not take into account the specific properties of posets, as opposed to general networks [10], or relied on supplementary statistical information whose presence we did not want to assume [7, 11]. Also, posets were structures about which we had few good intuitions, compared to more familiar structures such as trees, networks, or Euclidean spaces.

*Extended abstract prepared for the SIAM Workshop on Combinatorial Scientific Computing (CSC04).
†Corresponding author: Knowledge Systems and Computational Biology Team, 505-667-9096, `joslyn@lanl.gov`
‡Quantum and Classical Information Science Team, 505-667-0790, `smm@lanl.gov`

	Total Nodes	Leaves	Interior Nodes	Edges	Height	Width
Molecular Function	7.0K	5.6K	1.3K	8.1K	13	$\geq$ 3.5K
Biological Process	7.7K	4.1K	3.6K	11.8K	15	$\geq$ 2.9K
Cellular Location	1.3K	0.9K	0.4K	1.7K	13	$\geq$ 0.4K
GO	16.0K	10.6K	5.4K	21.5K	16	$\geq$ 5.9K

	Min	Max	Mean	Median
# Children/node	0	310	1.35	0.75
# Parents/node	0	6	1.35	1.71

Table 1: Summary statistics of the Gene Ontology, version `go_200309-assocdb.xml`.

Most significantly, compared to other combinatoric structures such as graphs, lattices, and trees, poset theory seemed to be relatively undeveloped (for example, the first textbook appeared in 2003 [12]). Most of the central questions we needed to answer, such as measures of "level" and "distance" in posets, were not obvious. While this was surprising, on the other hand the existence of such large-scale combinatorial structures as data objects, and the need to discover and measure such things about them, is relatively novel.

To date we have brought our work on categorization in the GO primarily to the computational biology community [5, 6], and now relish the opportunity to share our ideas and questions with the combinatorial scientific computing community. In particular, this talk will address the following:

- Our approach to the categorization task based on scoring functions which balance coversage and specificity by using pseudo-distance measures between comparable nodes.
- Recent positive validation achieved by comparing our results to Interpro [4] annotations.
- Potential generalizations of our pseudo-distance measures to involve interval-valued rank, height and width of intervals between bounds of noncomparable nodes, and Markov processes on poset intervals.
- Development of analytical methods for large posets (e.g. finding embedded sub-trees and sub-lattices).
- Measurement of coherence among different semantic orderings on the same node set.
- Weighting mechanisms to account for relative density and sparsity in portions of the ontology.
- Methods to regularize mappings and measure similarity among different ontologies (e.g. between GO and EC) in terms of order preservation and poset homomorphisms of both poset nodes and annotations.

References

[1] The Gene Ontology Consortium: (2000) "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, v. **25**:1, pp. 25-29

[2] *DARPA Agent Markup Language Homepage*, `http://www.daml.org/`

[3] *Enzyme Structures Database*, `http://www.biochem.ucl.ac.uk/bsm/enzymes/`

[4] *Interpro Home*, `http://www.ebi.ac.uk/interpro/`

[5] Joslyn, Cliff; Mniszewski, Susan; Andy Fulmer; and Gary Heaton: (2003) "Measures on Ontological Spaces of Biological Function", *Pacific Symposium on Biocompuating PSB 03*

[6] Joslyn, Cliff; Mniszewski, Susan; Andy Fulmer; and Gary Heaton: (2003) "Structural Classification in the Gene Ontology", in: *Proc. 6th Bio-Ontologies Workshop, Intelligent Systems for Molecular Biology (ISMB 03)*

[7] Lord, Phillip, R. Stevens, A. Brass, and C. Goble: (2003) "Investigating Semantic Similarity Measures Across the Gene Ontology: The Relationship Between Sequence and Annotation", *Bioinformatics*, 19(10):1275-83

[8] *Medical Subject Headings*, `http://www.nlm.nih.gov/mesh/meshhome.html`

[9] *OWL Web Ontology Language Overview*, `http://www.w3.org/TR/owl-features/`

[10] Rada, Roy; Mili, Hafedh; and Bicknell, E et al.: (1989) "Development and Application of a Metric on Semantic Nets", *IEEE Trans. on Systems, Man and Cybernetics*, v. **19**:1, pp. 17-30

[11] Resnik, Philip: (1999) "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems in Ambiguity in Natural Language", *J. Artificial Intelligence Research*, v. **11**, pp. 95-130

[12] Schröder, Bernd SW: (2003) *Ordered Sets*, Birkhauser, Boston