

Improved Order Theoretical Techniques for GO Functional Annotation

CA Joslyn, SM Mniszewski, KM Verspoor, and JD Cohn
Los Alamos National Laboratory

March 3, 2005

We are pursuing a research program in the application of finite order theory to the data analysis needs of bio-ontologies such as the Gene Ontology (GO)¹ [1]. In particular, we have developed the POSOC categorization tool [3] as a novel knowledge discovery technique in bio-ontologies. Categorization takes either a collection of genes or other properties annotated to ontology nodes, or just the nodes themselves, and finds the single node, or few nodes, which best summarize or categorize the position of the set in the overall ontology. We have further employed POSOC in a hybrid architecture which categorizes BLAST neighborhoods of sequences used in the CASP6² protein function competition to derive putative functional annotations of novel sequences [2, 4]. We have made a number of recent methodological advances in building our algorithmic capabilities for analyzing and manipulating ontologies such as the GO. In particular we have:

1. A new pseudo-distance measure based on discrete Markov processes on the GO.
2. A new ability to measure interval-valued rank in terms of vertical level in the GO.
3. And order theoretical measures of horizontal distance based on so-called “fence” measures.

In this poster, we present the results of bringing these new techniques to our automated functional annotation architecture in the following ways:

- We show a sensitivity analysis of the categorization algorithm to the POSOC input parameters, in particular the choice of pseudo-distance function, scoring function, and the free “specificity” parameter.
- This analysis reveals a surprising lack of sensitivity of the POSOC algorithm to the distance measure, despite very strong arguments to the contrary that this should be expected. We hypothesize that this is a function of the depths of the annotations used, and the relative lack of multiple inheritance (multiple children per node), especially at the lower levels of the GO. Quantitative confirmation of this hypothesis depends precisely on the new measures we have developed, and these are shown.
- Finally, prior to the categorization step in our architecture, we analyze the BLAST neighborhoods of known sequences with respect to the GO. In particular, we calculate the distribution of annotation terms within the GO with respect to height, depth, and vertical vs. horizontal spread in the GO, parameterized by BLAST evalue and annotation set.

References

- [1] Gene Ontology Consortium: (2000) “Gene Ontology: Tool For the Unification of Biology”, *Nature Genetics*, v. **25**:1, pp. 25-29
- [2] Cohn, Judith; Verspoor, K; Mniszewski, S; and CA Joslyn: (2004) “Predicting Protein Function Using Nearest Neighbor Categorization”, *Proc. 2nd Annual Rocky Mountain Regional Bioinformatics Conf. (Rocky 04)*
- [3] Joslyn, Cliff; Mniszewski, Susan; Fulmer, Andy; and GG Heaton: (2004) “The Gene Ontology Categorizer”, *Bioinformatics*, v. **20**:s1, pp. 169-177
- [4] Verspoor, Karin; Cohn, J; Mniszewski, SM; and CA Joslyn: (2004) “Nearest Neighbor Categorization for Function Prediction”, in: *Proc. 5th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP 05)*, in press

¹<http://www.geneontology.org>

²<http://predictioncenter.llnl.gov/casp6/Casp6.html>