

Investigating semantic similarity measures across the Gene Ontology: the relationship between interaction and annotation

Sonia Leach¹

Keywords: protein interaction, Gene Ontology, semantic similarity measures

1 Abstract.

Interactions between proteins can either be measured explicitly by various biological assays, or inferred based on properties of the proteins. We say that a pair of proteins has a *semantic* relationship if there is a collection of sources suggesting an interaction between them. We investigated the correlation between the sources of evidence of interaction and the similarity of Gene Ontology (GO) terms. We asked what was the level of GO similarity between terms assigned to a pair of proteins given that a particular technique suggested an interaction, as well as what was the likelihood of measuring an interaction given that a pair had a high GO similarity.

Computing the similarity between GO terms is complicated by the fact that the structure of the graph is not uniform – the distance of a term to the root of a taxonomy varies widely across terms and is not immediately indicative of the semantic preciseness of the term. Similarity of terms is not simply path distance. In a similar study which compared semantic similarity of GO annotations to sequence similarity, three *information theoretic* measures were proposed for measuring similarity of GO terms [6, 7]. All three measures, denoted **Resnik, Lin, and Jiang**, rely on the notion that less frequently used terms are more informative.

To compare the correlation between evidence of protein interaction and semantic similarity of GO terms, we first tabulated the number of sources that indicated an interaction, implicitly or explicitly, between every pair of proteins in the Yeast genome. As sources of interaction information, we used physical and genetic interaction databases such as DIP, BIND, and MIPS [11, 1, 9], as well as sources describing phylogenetic profiles [10], gene-fusion events [2], location analysis [5], growth phenotypes [3], cellular location [4], transcription factor sites [8], essentiality and various other MIPS catalogs [9]. We then calculated all GO pairwise similarity scores, using each of the three metrics.

To answer the first question, we examined the distribution of GO similarity scores given each type of interaction evidence, for example for yeast two-hybrid or phylogenetic profiles. We found that the Jiang metric was most discriminative, having high GO similarity for most types of interaction evidence and across all three GO taxonomies. Generally, the more implicit interaction information sources (such as location analysis, co-motif identification and gene fusion) were less correlated with high GO similarity than explicit sources such as yeast two-hybrid.

For the second question, we divided the range of values for each similarity metric into a number of bins and tabulated the number of pairs in a particular range which also had a particular number of pieces of interaction evidence. We found that the Jiang metric correlated most strongly with the presence of interaction evidence, across all three taxonomies.

¹Brown University and University of Colorado, Mail Stop 8303, P.O. Box 6511, Aurora, CO 80045.
E-mail: sml@cs.brown.edu

The Resnik measure exhibited particularly poor performance in both tests, in contrast to the results of the earlier study comparing sequence similarity and GO similarity where the Resnik measure was the most discriminatory [6, 7].

References

- [1] Bader, G. D. *et al.* 2003. BIND: the Biomolecular Interaction Network Database, *Nucleic Acids Res.* 31(1):248–50.
- [2] Enright, A. J. *et al.* 1999. Protein interaction maps for complete genomes based on gene fusion events, *Nature* 402:86–90.
- [3] Giaever, G. *et al.* 2002. Functional profiling of the *S. cerevisiae* genome, *Nature* 418:387–391.
- [4] Kumar, A. *et al.* 2002. Subcellular localization of the yeast proteome, *Genes and Development* 16:707–719.
- [5] Lee, T. I. *et al.* 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science* 298:799–804.
- [6] Lord, P. W. *et al.* 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics* 19(10):1275–1283.
- [7] Lord, P. W. *et al.* 2003. Semantic similarity measures as tools for exploring the gene ontology, *Pac Symp Biocomput* 601–612.
- [8] Matys, V. *et al.* 2003. TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Res* 31(1):374–378.
- [9] Mewes, H.W. *et al.* 2000. MIPS: a database for genomes and protein sequences, *Nucleic Acids Res.* 28:37–40.
- [10] Pellegrini, M. *et al.* 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *PNAS* 96:4285–4288.
- [11] Xenarios, I. *et al.* 2000. DIP: the database of interacting proteins, *Nucleic Acids Res.* 28:289–291.