Summary of Discussion for
Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation
by P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble
Bioinformatics 19(10) 1275–1283
http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/10/1275

presented by Christopher Maier for
INLS 279: Bioinformatics Research Review
2006-02-01

This article presents the interesting idea of leveraging ontological annotations (specifically, Gene Ontology (GO) annotations, though the methodology presented is generic) of bioinformatics resources to perform semantic queries over such data. The idea is that proteins (or genes, or pathways, or …) that are annotated similarly will have similar biological "meaning." The authors present a metric that takes advantage of the ontological structure of GO to capture some notion of "similarity" between database items annotated with GO terms. They validate their measure by showing a positive correlation with BLAST-derived similarity measurements, and proceed to demonstrate a prototype search engine that utilizes this similarity metric to retrieve semantically similar database entries for a query item. The group was in agreement that the idea certainly has merit and is worthy of investigation, but that this particular paper was not overly rigorous in presenting its arguments.

The first critique concerned the writing itself. Many items were vaguely conveyed, such as the notion of the "GO database" — are they referring to http://www.godatabase.org, and if so, why not explicitly say? — and the identity of the three Evidence Codes that predominated in SWISS-PROT-Human. There were several distracting copy errors as well. While this is not necessarily indicative of "bad science" *per se*, it does set the reader on edge from the beginning, and is quite unexpected from Oxford University Press.

More meaningfully, we felt that the authors glossed over many important methodological steps. The creation of the graphs in Figures 2, 3, and 4 is not detailed; the general idea is expressed, but specific details are omitted. What exactly do each of the points in these graphs represent? Why were there no regression fits for these plots? Why give only covariance, and not correlation coefficients?

The group recognized the need and utility of making simplifying assumptions for research such as this, particularly in initial exploratory work. However, a more thorough discussion of the costs and benefits of the various assumptions made would be very helpful. The idea of a minimum subsumer almost certainly discards potentially useful information concerning the connections between nodes, as does the decision to treat all connections equally, regardless of kind, essentially ignoring the semantics of "part-of" in favor of the more prevalent "is-a". Using unweighted connections can also obscure important details; a potential weighting scheme discussed by the group could incorporate

a measure of how many database items were annotated with a particular term, perhaps rewarding links to such terms. The group acknowledged the difficulties underlying the incorporation of this information into a similarity metric, but would have liked a more rigorous look at the consequences of their elision.

The authors state that the "Traceable Author Statement" (TAS) is the highest-value Evidence Code for GO, yet one group member with knowledge of annotation procedures pointed out that, in fact, TAS is *not* the gold standard of annotation evidence. Instead, direct experimental evidence, such as indicated by "Inferred from Direct Assay" (IDA), is preferred. Furthermore, the claim that since the "Inferred from Sequence Similarity" (ISS) code is so widespread in SWISS-PROT-Human, this supports the quality of the paper's thesis is certainly questionable; researchers generally assign the *lowest* confidence to the ISS code. This is not because scientists have no confidence in algorithms such as BLAST; far from it — they realize that such evidence alone is not sufficient to make any definitive conclusions, but that in the presence of other data it can be persuasive corroborating evidence. Additionally, the section concerning the authors' intention in investigating evidence codes was rather vague and ill-defined.

In the discussion of outliers, the authors make the claim that approximately half of the annotations in SWISS-PROT-Human are incorrect. These mis-annotations were all attributed to the Proteome Inc. group, to which annotation responsibilities were contracted in the initial period of database annotation. Over the years, the quality of these annotations as a whole has come into question. We wonder, given the dubious quality of these annotations, why were attempts not made to eliminate them from the experimental corpus?

The search tool prototype described toward the end of the article is intriguing, but of uncertain utility, at least at this early stage of development. As the "Molecular Function" sub-ontology had the highest correlation with BLAST searching, the similar items returned in the ontology search could have just as easily been returned using BLAST instead. The results returned for the "Cellular Component" search, being all integral membrane proteins, seem too general to be useful. The "Biological Process" results, being all components of vision-related processes, are interesting, but low (and identical) similarity scores raise questions about the appropriateness of the proposed similarity metric.

In summation, the group found the concept of semantic similarity measures utilizing ontology annotations to be intriguing, and definitely a promising area for research. The presentation of data in this article, however, was not as deep or thorough as we felt it should have been. Too many questions were raised over the course of the discussion that could not be answered from the paper.