Semantic similarity across the Gene Ontology: relating sequence and annotation

P.W.Lord, R.D. Stevens, A. Brass and C.A.Goble
Department of Computer Science
University of Manchester
Oxford Road
Manchester
M13 9PL
UK
p.lord@russet.org.uk
robert.stevens@cs.man.ac.uk
abrass@man.ac.uk
carole@cs.man.ac.uk

July 15, 2002

Bioinformatics resources are rich in knowledge. They hold data, often in the form of sequences, which are then annotated with the biological community's understanding about these entities. This knowledge is usually held as scientific natural language. In this form its human readable, and understandable. Although it can be accessed by computer applications, it is not easy to interpret computationally.

It is partly because of these problems that there has been a growing interest in ontologies within the bioinformatics community[Stevens et al., 2000]. In essence, ontologies provide a mechanism for capturing a community's view of a domain in a sharable form, that is both accessible by humans, and computationally amenable.

The Gene Ontology (GO) is becoming a *de facto* standard for the annotation of gene products. One of the claims made for the Gene Ontology is that it should allow improved querying of databases [The Gene Ontology Consortium, 2001]. Different resources queried with the same term should recover all and only entities conforming to that notion. One obvious way to query a database would be to ask for all proteins *semantically similar* to a query protein.

This notion of semantic similarity has been used in other areas. For instance, articles within PubMed are marked with terms from the Medical Subject Headings (MESH) terminology (see `http://www.nlm.nih.gov/mesh/meshhome.html`). The PubMed service (see `http://www.pubmed.gov`) offers a resource by which it is possible to retrieve related articles to the one in question. In essence, this is semantic similarity and is performed computationally via a series of lexical techniques [Wilbur and Yang, 1996].

We have investigated using an information content base measure of seman-

tic similarity, which is based on the notion that less frequently used terms are more informative. This idea is familiar from most text based search engines. Very commonly occurring words such as "the" and "a" are simply removed from queries, as they add no value to the search. Information content based measures were initially developed to operate over the WordNet dictionary/thesaurus [Fellbaum, 1998], for such purposes such as word sense disambiguation [Resnik, 1999], but these measures will work over any taxonomy, such as GO.

For WordNet these measures have been validated over small humanly generated data sets, and it was unclear whether they were appropriate, or biologically useful, when used over GO. In order to test this therefore, we used the notion that proteins with similar sequences, should be associated with similar annotation. By comparing the sequence similarity of two proteins, as determined by BLAST bit score, with the semantic similarity of the annotation, we found a highly significant correlation. More over this correlation was greatest with the semantic similarity calculated from molecular function aspect of GO, which fits well with biological expectations.

We present further analysis of the behaviour of these measures when limiting associations to specific evidence tags, and by including or ignoring links within GO. We also present prototype applications of these measures, which enable us to validate, or to query GO based ontological annotation. We believe this represents the first steps toward use of semantic similarity measures as a valuable tool in the armoury of the researcher.

References

[Fellbaum, 1998] Fellbaum, C., editor (1998). *WordNet:- An electronic lexical database*. MIT Press, Cambridge, Massachusetts.

[Resnik, 1999] Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

[Stevens et al., 2000] Stevens, R., Goble, C., and Bechhofer, S. (2000). Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics*, 1(4):398–416.

[The Gene Ontology Consortium, 2001] The Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–33.

[Wilbur and Yang, 1996] Wilbur, W. J. and Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med*, 26(3):209–22.