# Investigating semantic similarity measures across the Gene Ontology: the relationship between 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

## Abstract

**Motivation:** Many bioinformatics data resources not only hold data in the form of sequences, but also as annotation. In the majority of cases, annotation is written as scientific natural language: this is suitable for humans, but not particularly useful for machine processing. Ontologies offer a mechanism by which knowledge can be represented in a form capable of such processing. In this paper we investigate the use of ontological annotation to measure the similarities in knowledge content or "semantic similarity" between entries in a data resource. These allow a bioinformatician to perform a similarity measure over annotation in an analogous manner to those performed over sequences. A measure of semantic similarity for the knowledge component of bioinformatics resources should afford a biologist a new tool in their repetoire of analyses.
**Results:** We present the results from experiments that investigate the validity of using semantic similarity by comparison with sequence similarity. We show a simple extension that enables a semantic search of the knowledge held within sequence databases.
**Availability:** Software available from `http://www.russet.org.uk`
**Contact:** `p.lord@russet.org.uk`.

## 1 Introduction

Bioinformatics resources are rich in knowledge. They hold data, often in the form of sequences, which are then annotated with the community's understanding about those entities. This annotation or knowledge component of a resource is usually held in scientific natural language as text. In this form, it is human readable and understandable, but it is not easy to interpret computationally.

It is partly because of these problems that there has been growing interest in ontologies within the bioinformatics community (Stevens *et al.*, 2000). Ontologies provide a mechanism for capturing a community's view of a domain in a shareable form, that is both accessible by humans and computationally amenable. An ontology provides a set of vocabulary terms that label concepts in the domain. These terms should have definitions and be placed within a structure of relationships, the most important being the "is-a" relationship between *parent* and *child* and the "part-of" relationship between *part* and *whole* (Winston *et al.*, 1987; Odell, 1998). By capturing knowledge about a domain in a shareable and computationally accessible form, ontologies can provide defined, accessible and computable semantics about the domain knowledge they describe.

Currently, one of the most important ontologies within the bioinformatics community is the Gene Ontology (GO) (The Gene Ontology Consortium, 2001). GO comprises three orthogonal taxonomies or aspects, that hold terms that describe the attributes of *molecular function*, *biological process* and *cellular component* for a gene product. GO is a rapidly growing collection of about 11 000 phrases, representing terms or concepts, held within a Directed Acyclic Graph (DAG), part of which is shown in Figure 1. Terms can have multiple parents, as well as multiple children along the "is-a" relationships.

The terms held within this structure are used to annotate database entries (GO Consortium, 2002b). As they form a standard vocabulary across many biological resources such as SWISS-PROT (Bairoch & Apweiler, 2000), this shared understanding provides a valuable, computationally accessible form of the community's knowledge about these attributes. Information about the evidence for this knowledge is also provided by GO in the form of "Evidence Codes" (GO Consortium, 2002a). These codes are a simple controlled vocabulary that describe the nature of the evidence that is available to support a particular association.

One of the claims made for GO 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. The shared understanding should improve retrieval consistency across resources and the recall and precision within resources. One obvious alternative way to query a database would be to ask for 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 up with terms from the Medical Subject Headings (MeSH) terminology (MESH, 2002), which is a taxonomy of biomedical terms. The PubMed service (pubmed, 2002) 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 & Yang, 1996). Documents are similar if they have a similar content. This is measured by the words common to abstracts, words common to titles and MeSH terms in common. Words are weighted to indicate their importance in describing a document. This technique only uses the lexical content of MeSH, rather than any of its structure. Performing a search in Entrez, the search interface for PubMed, using only a MeSH term will, however, return documents marked-up with that term and any child term. This gives a small degree of semantic similarity, but uses no metric to judge the degree of similarity.

Bioinformaticians have realised that the computational use of the knowledge component is important. Similarity between annotation and literature has been shown to augment sequence similarity searches (Chang *et al.*, 2001; MacCallum *et al.*, 2000). These authors augmented PSI-BLAST (Altschul *et al.*, 1997) with similarity scores calculated over the annotations and Medline references cited by entries retrieved by the sequence similarity search. These were used to prune the results retrieved by each iteration to those most semantically similar to the query sequence. Both of these augmented PSI-BLAST's used the same statistical lexical approach developed for PubMed similarity.

In this paper we use an *information content* based measure of semantic similarity. This approach was originally developed using WordNet (Fellbaum, 1998), which is a computationally amenable dictionary/thesaurus, although to our knowledge such measures have not been previously applied to GO. Unlike lexical approaches used on MeSH terms, this measure makes explicit use of the ontological structure. We describe a series of investigations which explore the validity of this measure when applied to GO.

## 2 Semantic Similarity Measures

Clearly, if two proteins are both annotated as "transmembrane receptor", (GO:0004888) they have a similar semantic description of their function. If one were annotated, less precisely, as just "receptor", (GO:0004872) then they have a slightly less similar function than before and are correspondingly semantically less similar.

Various measures have been developed for quantifying this notion of semantic similarity. Early techniques have used path distances between terms (Rada *et al.*, 1989). One of the main difficulties with this approach is that it assumes that all of the semantic links are of equal weight, which appears to be a poor assumption. For example, the pair "photoreceptor", (GO:0009881) and "transmembrane receptor", (GO:0004888) are semantically more closely related than "chaperone", (GO:0003754) and "signal transducer", (GO:0004871). Inspection of Figure 1 reveals these two pairs would have identical similarities as they have an immediate common parent, but the former would appear to be more closely related, than the latter.

There are a number of ways that edges could be weighted. Generally, the greater the distance from the root of the graph, the more specific the terms. However GO varies widely in the distance of nodes from the root. So, "high-affinity tryptophan transporter", (GO:0005300) is 14 terms deep, while "anticoagulant", (GO:0008435) is only 3 terms deep, and not significantly less semantically precise. It would appear that the depth of GO reflects mostly the vagaries of biological knowledge, rather than anything intrinsic about the terms.

Instead of attempting to define similarity simply on the basis of the structure of the ontology, it is also possible to examine the usage of terms within the corpus (Resnik, 1999). This uses the notion of "information content". For instance, "chaperone", (GO:0003754) is a more informative term than "signal transducer", (GO:0004871), because the former is used several hundred times, while the latter is used several thousand times. This notion is familiar from most internet search engines. Searching with "alpha mating factor" may give information about yeast cells, while "sex pheromone" is likely to reveal a very different sort of biological information. The phrase "alpha mating factor" is more informative, because it occurs less often. With GO annotations, we can exploit the

usage of terms in the corpus to give a measure of information content.

In the case of GO we can also exploit the semantic links in the calculation of the information content for each concept. If the term "receptor", (GO:0004872) occurs, then implicitly, the concept "signal transducer", (GO:0004871) and "molecular function", (GO:0003674) have also occurred, as well as any other terms which subsume it. Generally, for semantic similarity, only the "is-a" links are considered (Resnik, 1999), although other semantic links can also be used.

In Figure 1 these probabilities are shown diagrammatically. In this case we have used the SWISS-PROT-Human proteins, and counted the number of times each concept occurs. A concept occurs if a term, or any of its children occur. The probability, $p(c)$, for each node is this value, divided by the number of times any term occurs. We can therefore guarantee that the probability for each node increases as we move up the graph toward the root, and that the probability for the root node occurring will be 1 (although the existence of "orphan terms" would invalidate this, see Section 3.1).

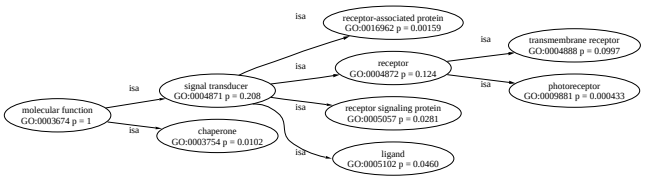


**Fig. 1.** Probabilities in the Gene Ontology. Each node is annotated with its GO accession and the probability of this term occurring in the SWISS-PROT-Human database. See Section 2 for details. This figure was produced from GO, using the graphviz tools (`http://www.graphviz.org`).

Once we have calculated these probabilities, there are a variety of different mechanisms for calculating the semantic similarity between terms (Jiang & Conrath, 1998; Lin, 1998). In this paper we have used the simplest of these measures (Resnik, 1999). This measure is based on the information content of shared parents of the two terms, as defined in Equation (1), where $S(c1, c2)$ is the set of parental concepts shared by both $c1$ and $c2$. As GO allows multiple parents for each concept, two terms can share parents by multiple paths. We take the minimum $p(c)$, where there is more than one shared parent. We call this $p_{ms}$, for *probability of the minimum subsumer*,

$$p_{ms}(c1, c2) = \min_{c \in S(c1,c2)} \{p(c)\} \quad (1)$$

The similarity score between two terms is then given by Equation (2).

$$\mathrm{sim}(c1, c2) = -\ln p_{ms}(c1, c2) \quad (2)$$

# 3 Validating Semantic Similarity

We can create a measure of semantic similarity, but how do we validate such a measure? SWISS-PROT and other resources now have conceptual annotations from GO and thus we have the knowledge, together with the sequence it describes. One of the tenets of biology is that a protein's sequence relates to its function. So highly similar sequences should be highly semantically similar. Taking protein sequences in pairs and plotting sequence similarity against semantic similarity should show a relationship. We used this hypothesis to test our measure. We next explored the other GO taxonomies of biological process and cellular component. Later in our experiments we looked at the use of evidence codes in annotations and aspects of the structure of GO and its influence on our scores.

## 3.1 Adapting the Similarity Measures to GO and SWISS-PROT

One feature of GO is that when a term is "part-of" another term, it often has no "is-a" link. This is deliberate: to reduce the number of abstract terms, such as "ribosomal component" (which would subsume terms such as "small ribosomal subunits", (GO:0015359)), which were not wanted for the annotation task for which GO was designed (M.Ashburner pers.comm.). Logically, of course, all terms must be a kind of another term. These *orphan* terms within GO need to be provided with links for the purposes of our investigation. We simply linked them directly to the root of their taxonomy. This is perhaps semantically impoverished (for example, a "granum", (GO:0009542) becomes a kind of "cellular component", (GO:0005575), rather than a kind of a "chloroplast component"), but this ontological sleight of hand made our semantic measurement possible.

It is also unclear how we should address the different link types. Except where stated explicitly (see Section 4.3), we consider the links equally. We took this approach because in GO there is a bias in link type usage between the different sub-ontologies (molecular function, 6207 is-a's to 35 part-of's, cellular component, 542 to 619, biological process, 5697 to 989). The semantic impoverishment would, therefore, have been very different between these different ontologies, making meaningful comparisons difficult. Conversely it reduces the problem of orphan nodes, which only occur when is-a's links alone are considered.

In this paper, we are mostly interested in the semantic similarity between proteins, rather than GO terms *per se*. We therefore need a method for combining these measures as proteins may be annotated with more than a single term. In previous work, based on WordNet, a similar problem has been found, as individual words have more than one sense (Resnik,

1999). In this case, the semantic similarity between words was calculated by simply taking the maximum similarity between any word sense, as only one sense of a word is used at a time. With GO annotated gene products, this is not the case, rather the gene product will have all of the roles attributed to it by annotators, using GO, at the same time. We have therefore taken the average similarity between all terms. In practise within SWISS-PROT-Human, especially when considering only "traceable author statement" associations (which, except where explicitly stated, has been the case in this paper), most proteins have been annotated with only a single GO term from each aspect (for "molecular function", 2929 single annotations, compared to 863 with two or more).

All the analysis in this paper was performed using a library generated for the purpose. This library is freely downloadable, and full details are published elsehwere (Lord *et al.*, 2003).

# 4 Investigating Semantic and Sequence Similarity

Previous work on semantic similarity had defined similarity measures either with specific applications in mind, such as malapropism detection, or word sense disambiguation (see (Fellbaum, 1998) and references therein), and had tested results against the expectations of people (Resnik, 1999; Budanitsky & Hirst, 2001). The difficulty in these cases is that such human generated test sets are often very small, a problem which is exacerbated in our case as biological experts are rarer than those with a working knowledge of English.

In order to overcome this difficulty we wished to validate our semantic similarity measures against some other metric. We used the relationship between sequence and annotated function as a means of validating our measure.

We therefore wished to obtain a set of protein pairs with varying degrees of sequence similarity. The standard BLAST tool provides just this by returning a ranked set of sequences similar to a query sequence. We have chosen to use the "Bit Score" as a mesure of sequence similarity, as this is independent of database size.

## 4.1 Comparing Semantic Similarity Across GO Aspects

The results, shown in Figure 2, show that there is a good correlation between sequence similarity and semantic similarity. This correlation is greater when measured against the "molecular function" aspect. There is still a correlation with the other two aspects, particularly at higher sequence and semantic similarity levels. This is unsurprising. As sequence similarity increases, so does the chance that these proteins are homologues, in which case they are likely to be identically annotated for all aspects.

It therefore appears that the semantic similarity correlates, as expected, when measured against a standard sequence similarity measure. This therefore serves as a good validation of the semantic similarity measure:- we find the results predicted from our understanding of biology.

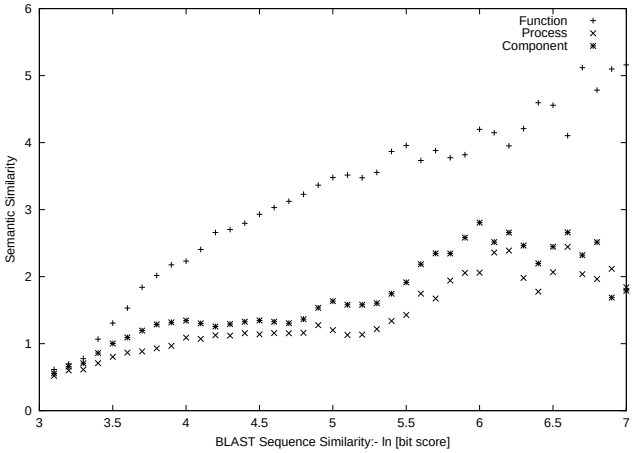


**Fig. 2.** Comparing sequence and semantic similarity. BLAST searches were performed for each SWISS-PROT-Human protein, and all matches analysed for semantic similarity with the search protein. For "function" $n = 68142$, covariance = 0.58, "process" $n = 76089$, covariance = 0.28, "component" $n = 39394$, covariance = 0.36.

## 4.2 The Relationship Between Semantic Similarity and Evidence Codes

Initially we were interested in the usage of evidence codes within SWISS-PROT-Human, and in general in the database. These inform us as to how the annotation was made: we would, for example, wish to exclude those proteins whose annotation is based purely upon sequence similarity. It appears that only three of the codes are in common usage, at least within SWISS-PROT-Human. Further analysis was therefore performed on data with only these codes.

Of the three commonly used evidence codes, "Traceable Author Statement" (TAS) is generally regarded as the highest standard of evidence. It is assigned where evidence is found in primary literature. GO associations assigned this evidence code might be expected to be the most accurate. The high percentage of these associations (70%, compared to 30% for the GO database as a whole), was one of the more important reasons for the choice of SWISS-PROT-Human within this work, and also the reason why only TAS associations were used in other parts of this work.

We therefore examined semantic similarity measurements considering GO annotations assigned the various evidence codes. This was limited to the functional aspect of GO, as this showed the most marked correlation with sequence similarity.

As shown in Figure 3 all the semantic similarity measurements against the three GO aspects show a correlation with sequence similarity. However when only TAS GO annotations are considered, the correlation is much greater.

Within the GO database as a whole, other evidence codes, particularly ISS or "Inferred from Sequence Similarity" are much more widely used. Given the validity of the relationship between semantic and sequence similarity, we can consider this to be a measure of the quality of the evidence. It would be of great interest, therefore, to extend the analysis to the whole GO database, as this might suggest which of the various evidence codes are most reliable.

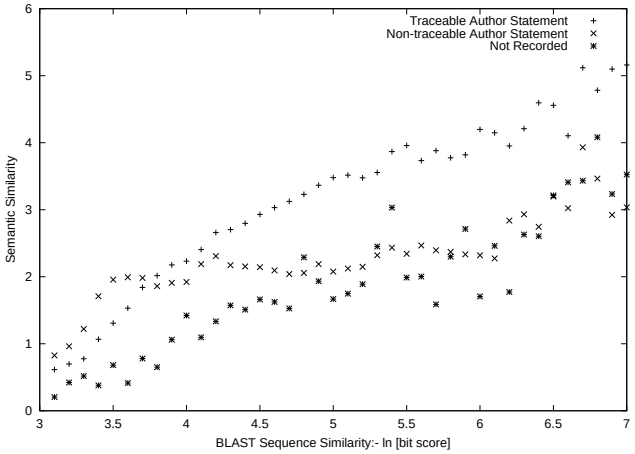


**Fig. 3.** Semantic similarity over the molecular function aspect and evidence codes. Semantic similarity scores were calculated on the basis of associations with the shown Evidence Code. The probability scores described in Figure 1 were calculated using only associations with the given Evidence Code. For "TAS", $n = 68142$, covariance = 0.58, "NAS", $n = 19631$, covariance = 0.26, "NR" $n = 2601$, covariance = 0.49.

## 4.3 Effect of Using Semantic Links in Semantic Similarity

One of the main differences between GO and a simple controlled vocabulary, such as the SWISS-PROT keywords, are the existence of explicit relationships between the different terms. The semantic similarity measures described in this paper make explicit use of this information. Does the inclusion of these semantic relationships actually provide useful information?

With the semantic similarity measure described, we can ignore all of this link information, effectively turning each term into an orphan term (see Section 3.1). Ignoring links changes the structure of GO from a heavily connected graph, to a simpler one where each term inherits directly and only from the root term: Essentially a set of terms akin to SWISS-PROT keywords. Alternatively, we can consider only links of a single type, either "is-a" or "part-of".

We investigated semantic measures either using all the link information, just "is-a" links, or no links at all. The results for the "molecular function" ontology are shown in Figure 4. Very little difference can be seen between graphs using all links, or just "is-a" links. This is to be expected, as for this aspect of GO almost all links are of the "is-a" type (6167 out of 6202). If no links are included the semantic similarity drops markedly, particularly in the middle part of this graph. At moderate levels of sequence similarity, proteins will often share similar, but not identical GO annotations. Consequently, these terms will only contribute to our semantic similarity measure if the links are included. Conversly, where sequence similarity is very high, GO annotations may well be identical, so ignoring links makes little difference. It appears that our semantic similarity measures are improved by the usage of the link information, which therefore provides a significant advantage over the use of a pure controlled vocabulary.

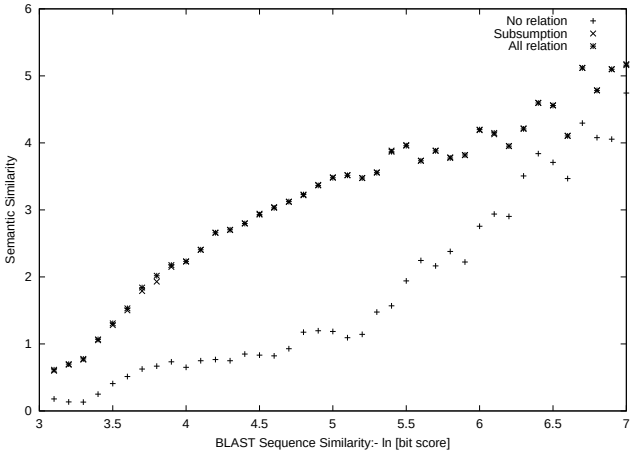


**Fig. 4.** Semantic similarity over the molecular function aspect and semantic relationships. BLAST searches were performed and analysed as in Figure 2. Term probabilities and semantic similarities were calculated using none, is-a or all semantic relationships. For "all" $n = 68142$, covariance = 0.58, for "subsumption" $n = 68142$, covariance = 0.58, for "none" $n = 68142$, covariance = 0.38.

### 4.4 Investigating Outliers Between Semantic and Sequence Similarity

Although we have shown a strong correlation between semantic and sequence similarity, there were a number of protein pairs which did not obey this trend. In particular we were interested in those proteins which showed very high semantic similarity but little sequence similarity. We therefore analysed those protein pairs with low sequence similarity and high semantic similarity.

There appear to be several categories of protein pairs in this area:

- "polymorphic" groups, where there are two or more classes of protein involved in the same process. See Table I. This group includes pairs, some of which heterodimerize, or are identified as sub-families by the various protein family databases.

- Hyper variable protein families. See Table II. The distinction between this and the last category is somewhat arbitrary, but we have applied it where sub-families are not refered to in the protein family databases.

- Mis-annotations. About half of the proteins appear to be incorrectly annotated (Table III). In most cases it is clear how this annotation has occurred. There are several cases, which are annotated in SWISS-PROT as being "$x$-like" but have been annotated in GO as "$x$". Others appear to be "spelling mistakes". So a spermine synthase is annotated as a "spermidine synthase", (GO:0004766). All of the mis-annotations reported here stem from the dataset incorporated from manual GO annotation by Proteome Inc., and extracted via LocusLink (E. Camon. pers.comm.).

For all of those protein pairs which identified a mis-annotation, the correction of these errors would lessen the semantic similarity scores (data not shown), which would, in turn, make them more reflective of the trend. It would be predicted therefore that as the use of GO improves and becomes more accurate, the correlation should strengthen. It would also appear that semantic similarity measurements could form a valuable tool for those seeking to check the annotations of proteins with GO terms.

Additionally we were interested in protein pairs with very low semantic similarity, but very high sequence similarity. In this section of the graph generally one or both of the proteins are "under-annotated". By this we mean that a fairly general term has been used when a more specific term would be better. There appear to be three main reasons for this; the lack of biological knowledge, the lack of a more specific GO term, or mis-annotations (data not shown).

## 5 Semantic Searching of GO Annotated Resources

Although we have been using sequence similarity in an attempt to validate semantic similarity, it also raises the obvious question of whether it is possible and useful to provide a search tool analogous to BLAST, which directly answers the question of whether there are any semantically similar proteins to a query protein or other biological entity annotated with GO terms.

We developed a search tool which tests a given query protein against all the others in SWISS-PROT-Human, and generates a ranked list of semantically similar proteins. Results for a sample protein are shown in Table IV. We have separated out lists from the different aspects of GO.

| Swissprot ID | Description | Similarity |
|---|---|---|
| a) Molecular Function | | |
| OPSG_HUMAN | Green-sensitive opsin (Green cone photoreceptor pigment). | 8.15 |
| OPN4_HUMAN | Opsin 4 (Melanopsin). | 7.23 |
| OPSB_HUMAN | Blue-sensitive opsin (Blue cone photoreceptor pigment). | 4.92 |
| 5H6_HUMAN | 5-hydroxytryptamine 6 receptor (Serotonin receptor) | 3.92 |
| A1AA_HUMAN | Alpha-1A adrenergic receptor (Alpha 1A-adrenoceptor) | 3.92 |
| A1AB_HUMAN | Alpha-1B adrenergic receptor (Alpha 1B-adrenoceptor). | 3.92 |
| b) Biological Process | | |
| AIPL_HUMAN | Aryl-hydrocarbon interacting protein-like 1. | 2.89 |
| CNCG_HUMAN | Retinal cone rhodopsin-sensitive cGMP | 2.89 |
| CNRA_HUMAN | Rod cGMP-specific 3',5'-cyclic phosphodiesterase | 2.89 |
| CNRC_HUMAN | Cone cGMP-specific 3',5'-cyclic phosphodiesterase | 2.89 |
| CNRD_HUMAN | Retinal rod rhodopsin-sensitive cGMP | 2.89 |
| CRB1_HUMAN | Beta crystallin B1. | 2.89 |
| c) Cellular Component | | |
| 1A01_HUMAN | HLA class I histocompatibility antigen | 1.86 |
| 5H1A_HUMAN | 5-hydroxytryptamine 1A receptor (5-HT-1A) | 1.86 |
| A1A2_HUMAN | Sodium/potassium-transporting ATPase alpha-2 chain | 1.86 |
| A1AA_HUMAN | Alpha- 1A adrenergic receptor | 1.86 |
| A33_HUMAN | Cell surface A33 antigen precursor | 1.86 |
| ACHA_HUMAN | Acetylcholine receptor protein | 1.86 |

**Table IV.** The table shows the results of a search over SWISS-PROT-Human, using the "OPSR_HUMAN" (accession no. P04000) protein as a query. Semantic similarities have been calculated for the three GO aspects using associations with any evidence code, and any semantic links. Results have been elided to show illustrative examples.

In this case we have searched with the protein "OPSR_HUMAN" (Red sensitive Opsin) (accession no. P04000). As might be expected from the molecular function aspect, a number of similar and related proteins are retrieved. As would be predicted from the results described in Section 4.1, this list is similar to that which would be retrieved using a BLAST search.

Results from the other aspects, however, are different. The biological process aspect has retrieved a variety of different proteins, with very different sequences, which are all however involved in vision, while the cellular component aspect retrieves other integral membrane proteins.

This suggests that the semantic similarity measure can be used to usefully retrieve related proteins from a database. It offers alternative dimensions along which to search. All three aspects of GO are useful for this task, returning a different, but equally valuable view on the protein.

| Protein A (ID) | Description | Protein B (ID) | Description | Seq. Sim. | Sem. Sim | Notes |
|---|---|---|---|---|---|---|
| DFFA_HUMAN | DNA fragmentation factor alpha subunit | DFFB_HUMAN | DNA fragmentation factor 40 kDa subunit | 3.49 | 7.79 | Heterodimers. |
| TKN1_HUMAN | Protachykinin 1 [Precursor] | TKNK_HUMAN | Neurokinin B [Precursor] | 3.23 | 7.79 | Sub-families. |
| LCFB_HUMAN | Long-chain-fatty-acid–CoA ligase 2 | VLCS_HUMAN | Very-long-chain acyl-CoA synthetase | 3.52 | 7.39 | Sub-families. |

**Table I.** The table shows protein pairs which heterodimerise, or which have been identified as members of sub-families in one or more protein family databases (data not shown.).

| Protein A (ID) | Description | Protein B (ID) | Description | Seq. Sim. | Sem. Sim | Notes |
|---|---|---|---|---|---|---|
| AKA5_HUMAN | A-kinase anchor protein 5 | AKAC_HUMAN | A-kinase anchor protein 12 | 3.74 | 7.10 | |
| CTR4_HUMAN | Cationic amino acid transporter-4 (CAT-4) | YLA1_HUMAN | Y+L amino acid transporter 1 | 3.49 | 7.10 | |
| EGF_HUMAN | Pro-epidermal growth factor precursor (EGF) | EREG_HUMAN | Epiregulin precursor | 3.85 | 7.10 | |
| FABE_HUMAN | Fatty acid-binding protein, epidermal (E-FABP) | FABI_HUMAN | Fatty acid-binding protein, intestinal (I-FABP) | 3.99 | 7.10 | |
| GBG3_HUMAN | Guanine nucleotide-binding protein | GBGB_HUMAN | Guanine nucleotide-binding protein | 3.85 | 7.39 | G proteins. |
| HMGC_HUMAN | High mobility group protein HMGI-C | HMGI_HUMAN | High mobility group protein HMG-I/HMG-Y | 3.35 | 7.79 | AT binding. |
| IPKA_HUMAN | cAMP-dependent protein kinase inhibitor, alpha form | IPKG_HUMAN | cAMP-dependent protein kinase inhibitor, gamma form | 3.82 | 7.79 | |
| PE21_HUMAN | Prostaglandin E2 receptor, EP1 subtype | PE22_HUMAN | Prostaglandin E2 receptor, EP2 subtype | 3.78 | 7.10 | |

**Table II.** The table shows protein pairs which, although are "outliers" appear to have been annotated correctly, and therefore represent highly variable families.

| Protein A (ID) | Description | Protein B (ID) | Description | Seq. Sim. | Sem. Sim | Notes |
|---|---|---|---|---|---|---|
| SPEE_HUMAN | Spermidine synthase (EC 2.5.1.16) | SPSY_HUMAN | Spermine synthase (EC 2.5.1.22) | 3.97 | 7.79 | The latter is mis-annotated as a spermidine synthase, when in fact its spermine synthase. [1] |
| THI2_HUMAN | Mitochondrial thioredoxin precursor (MT-TRX). | TXNL_HUMAN | Thioredoxin-like protein | 3.75 | 7.39 | Annotated with an obsolete term. TXNL_HUMAN is only thioreduxin. [1] |
| INL3_HUMAN | Leydig insulin-like peptide precursor | INS_HUMAN | Insulin precursor. | 3.44 | 7.79 | Annotated as insulin, although the former is only "insulin-like". [1] |
| DNM1_HUMAN | DNA (cytosine-5)-methyltransferase 1 | DNM2_HUMAN | DNA (cytosine-5)-methyltransferase-like protein 2 | 3.98 | 7.39 | Annotated as methyltransferases, although the latter is not. [1] |
| PTHR_HUMAN | Parathyroid hormone-related protein precursor | PTHY_HUMAN | Parathyroid hormone precursor | 3.50 | 7.79 | Annotated as "cAMP generating". Problem with GO structure [2]. |
| ZO2_HUMAN | Tight junction protein ZO-2 | CSKP_HUMAN | Peripheral plasma membrane protein CASK | 3.89 | 7.10 | Annotated as "membrane-associated protein with guanylate kinase activity", (GO:0004384). Problem with GO structure [2]. |

**Table III.** Incorrect GO annotations. The table shows SWISS-PROT-Human associations which appear to be incorrect. See further details in text. [1] These are confirmed to be incorrect annotations, that were incorporated into SWISS-PROT human GOA file using the manual GO annotation of Proteome Inc. extracted via Locus Link. (E. Camon pers.comm. and (Camon *et al.*, 2002). [2] These result from errors in the GO structure, as confirmed by the GO editors (M.Harris, M. Ashburner, pers.comm.)

These data also show one of the problems with this sort of search tool. Many of the results returned have identical similarity values, therefore requiring a second ranking mechanism (the current search tool uses alphabetic ordering of the Swissprot ID, which is clearly less than satisfactory). This problem stems from two sources. Firstly, the relatively small size of GO. So all the proteins in Table IV c) have been retrieved through the term "integral plasma membrane protein", (GO:0005887). Clearly this problem should lessen as GO increases in size and coverage. Secondly, the similarity measure used, which considers only the information content of shared parents of the query terms, $p_{ms}$, as defined in Equation (1) meaning that the semantic distance between many different GO terms is identical. It may be that other measures, which also use the information content of query terms (Jiang & Conrath, 1998; Lin, 1998), may help to ameliorate this problem. In conclusion, we believe that even our primitive search tool is already useful.

# 6 Discussion

In this paper we have investigated semantic similarity measures, and their application to ontological annotations of the SWISS-PROT database. Instead of sequence similarity, we are asking "is what we know about these proteins similar?". In all cases semantic similarity is correlated with sequence similarity, but this correlation is more marked against the molecular function aspect, which we would predict from our understanding of biology.

Having provided initial validation of these similarity measures, we have also investigated the use of evidence codes within GO, and semantic similarity measures using only associations with given evidence codes. This suggests that on a large scale statistical basis the associations with "Traceable Author Statement" evidence are the most informative.

We have also investigated the use of the ontological structure and how this affects the similarity measure, by "flattening" GO into a pure controlled vocabulary, and shown that they provide important information. It should be noted that

as GO increases in size the relationships are likely to get more important, as the chance that any two proteins will share an identical GO term will decrease. The semantic similarity measure should avoid a well known problem with a controlled vocabulary; how large should the vocabulary be? If it is too small its not expressive enough, too large then it becomes free text or simply unmanageable. Semantic similarity measurements across GO should continue to work as GO expands, indeed, they should improve.

Future work will explore the effects of the different semantic links in ontologies. Currently, all links are treated as "is-a" links: throwing away semantic information, and how they could contribute differently to semantic similarity needs to be addressed.

Two direct applications of this measure have been developed, checking for errors during the annotation process, and a search tool. Although both tools need further work before being useful as an end user tool, they serve as a proof of concept. A large number of potential uses for semantic similarity measures have been considered. By allowing ranking of GO terms, they should support the original intention of GO, to provide a unifying force between different, and often heterogeneous, databases. The current study has focused mainly on the molecular function aspect of GO. It would be of great interest to investigate the relationships between semantic similarity and co-expression as revealed by microarray experiments. It be expected that the biological process aspect would be of great use in this context.

Resource annotation and the bio-medical literature have been recognised as a valuable resource in performing sequence analyses (Chang *et al.*, 2001; MacCallum *et al.*, 2000). These approaches have used a statistical, lexical approach to comparisons of the knowledge component. This paper has presented a metric for semantic similarity based upon ontological annotation of resources. Such annotations are likely to spread, offering a widespread, alternative mechanism for exploring and validating bioinformatics knowledge, and providing the basis for valuable tools for the Conceptual Biologist (Blagosklonny & Pardee, 2002).

**Acknowledgements:** The authors would like to thank the authors of the freely available libraries, in particular, the GO database and API, bioperl, and EMBOSS who made this work possible. Thanks are also due to E.Camon, M.Harris, and M. Ashburner, both for the comments on the manuscript, and on some of the data shown. This work was funded under the EPSRC/BBSRC Bioinformatics Programme (Grant number: BIF/10507)

# References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* **25** (17), 3389–402.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res,* **28** (1), 45–8.

Blagosklonny, M. V. & Pardee, A. B. (2002). Unearthing the gems. *Nature,* **416**, 373.

Budanitsky, A. & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics* , Pittsburgh:.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., & Apweiler, R. (2002). The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL and InterPro. *Genome Research,* . Submitted.

Chang, J., Raychaudhuri, S., & Altman, R. (2001). Including Biological Literature Improves Homology Search. *Pacific Symposium on Biocomputing,* **6**, 374–383.

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

GO Consortium (2002a). `http://www.geneontology.org/doc/GO.Evidence.html`.

GO Consortium (2002b). `http://www.geneontology.org/goa`.

Jiang, J. J. & Conrath, D. W. (1998). Semantic similarity based on corpus statistics and lexical taxonomy. In: *Proceedings of International Conference on Research in Computational Linguistics* , Taiwan: ROCLING X.

Lin, D. (1998). An information-heoretic definition of similarity. In: *Proc. 15th International Conf. on Machine Learning* pp. 296–304, Morgan Kaufmann, San Francisco, CA.

Lord, P., Stevens, R., Brass, A., & Goble, C. (2003). Semantic similarity measures as tools for exploring the Gene Ontology. In: *Pacific Symposium on Biocomputing* volume 8 pp. 601–612,.

MacCallum, R. M., Kelley, L. A., & Sternberg, M. J. (2000). SAWTED: structure assignment with text description–enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics,* **16** (2), 125–9.

MESH (2002). `http://www.nlm.nih.gov/mesh/meshhome.html`.

Odell, J. (1998). Six Different Kinds of Aggregation. In: *Advanced Object-Oriented Analysis and Design Using UML* pp. 139–149, Cambridge University Press.

pubmed (2002). `http://www.pubmed.gov`.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics,* **1** (19), 17–30.

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, R., Goble, C., & Bechhofer, S. (2000). Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics*, **1** (4), 398–416.

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

Wilbur, W. J. & 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.

Winston, M., Chaffin, R., & Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. *Cognitive Science*, **11**, 417–444.